Adaptive Latent Diffusion Model for 3D Medical Image to Image Translation: Multi-modal Magnetic Resonance Imaging Study

Jonghun Kim, Hyunjin Park †
Department of Electrical and Computer Engineering
Sungkyunkwan University, Suwon, Korea
iproj2@skku.edu, hyunjinp@skku.edu

Abstract

Multi-modal images play a crucial role in comprehensive evaluations in medical image analysis providing complementary information for identifying clinically important biomarkers. However, in clinical practice, acquiring multiple modalities can be challenging due to reasons such as scan cost, limited scan time, and safety considerations. In this paper, we propose a model based on the latent diffusion model (LDM) that leverages switchable blocks for image-to-image translation in 3D medical images without patch cropping. The 3D LDM combined with conditioning using the target modality allows generating high-quality target modality in 3D overcoming the shortcoming of the missing out-of-slice information in 2D generation methods. The switchable block, noted as multiple switchable spatially adaptive normalization (MS-SPADE), dynamically transforms source latents to the desired style of the target latents to help with the diffusion process. The MS-SPADE block allows us to have one single model to tackle many translation tasks of one source modality to various targets removing the need for many translation models for different scenarios. Our model exhibited successful image synthesis across different source-target modality scenarios and surpassed other models in quantitative evaluations tested on multi-modal brain magnetic resonance imaging datasets of four different modalities and an independent IXI dataset. Our model demonstrated successful image synthesis across various modalities even allowing for one-to-many modality translations. Furthermore, it outperformed other one-to-one translation models in quantitative evaluations. Our code is available at https://github.com/jongdory/ALDM/

Figure 1. Overview of the proposed image-to-image translation process based on latent diffusion model. Our model utilizes the proposed MS-SPADE to transform the latent representation into the target latents and enables image synthesis in the desired target modality through conditioning.

1. Introduction

Multi-modal images are crucial for a comprehensive evaluation in medical image analysis. To identify clinically important biomarkers, multi-modal 3D images with complementary information are often required. Multi-modal imaging serves as a method to compensate for the limitations of individual imaging techniques and enables precise examination [25, 36, 38]. Medical image analysis, especially diagnosis of brain tumors, is typically performed using magnetic resonance imaging (MRI) and various MRI techniques such as 3D T1-weighted, T1 contrast-enhanced weighted, T2-weighted, and FLAIR are used, because the distinct modalities (T1, FLAIR, etc.) are different imaging sequences that provide varying contrasts and provide complementary information (e.g., edema, enhancing tumor, and necrosis/non-enhancing tumor) about brain tumor that appears over multiple consecutive slices requiring 3D approach. [2, 3, 27]. T1 is sensitive to acute bleeding and clearly displays contrast agents, helping identify vascular structures, tumors, and inflammation areas. T2 highlight conditions like edema or inflammation. FLAIR suppresses cerebrospinal fluid signals and is useful for detecting pathologies such as multiple sclerosis lesions [3, 27, 29, 37]. This is an important application of computer vision in medicine. However, in clinical settings, obtaining multiple modalities can be challenging due to factors such as scan costs, limited scan time, and safety considerations. Consequently, certain modalities may be missing. This absence of
modalities can have a detrimental impact on the quality of diagnosis and treatment. Furthermore, deep learning models that rely on multi-modal also suffer from reduced performance when crucial modalities are unavailable for training.

Many effective deep learning methods have been proposed to address the information gap caused by missing modalities [6, 11, 16, 23, 24, 40, 44]. One approach is to train a single model that learns the shared representation of multi-modalities to handle various missing scenarios [6, 24, 40, 44]. For example, with three out of four MRI modalities missing, we must consider 12 (3 × 4) missing scenarios. The number of scenarios grows quickly as the number of modalities grows. However, these methods, limited to specific tasks, struggle to generalize across diverse scenarios, not offering a fundamental solution.

Another approach is to generate missing modality images using generative models. Among the existing methods for generating missing modalities, generative adversarial models (GANs) [7, 11, 15, 19, 23, 46] have been used for medical image-to-image translation. However, these methods are 2D and thus do not consider the information from adjacent slices in 3D medical images. They generate slices independently making it difficult to ensure consistent information between adjacent slices. 3D methods for handling medical imaging data are computationally expensive and challenging to apply. To address this issue, a common approach is to train the model using patches during the training stage and generate results using a sliding window method during the inference stage. However, these methods may result in the loss of global information and a potential decrease in image translation performance. Furthermore, existing methods typically generate a given target modality from one or many source(s) (i.e., one-to-one or many-to-one) [11, 23, 45]. To generate multiple targets from one source modality, these methods need a separate model for each target, increasing complexity. Therefore, there is a need for a model that can perform image translation using 3D medical imaging without patch cropping and allow generating flexible targets from a single source modality.

In recent years, the diffusion model (DM) has emerged as a powerful generative model with high-quality generation capabilities, positioning itself as a potential alternative to GANs [9, 12, 17, 31, 34, 35]. The diffusion model leverages cross-attention and flexible conditioning to enable the generation of desired images. Additionally, the latent diffusion model (LDM) defines a denoising task in the compressed latent space of an image significantly reducing computational costs while demonstrating the ability to generate high-quality images at a lower cost [34]. LDM demonstrated its applicability to 3D medical images without the need for patch cropping. This advancement allows for the application of LDM to medical images even in their native 3D form [21, 31].

In this study, we propose a model that utilizes the conditioning mechanism of LDM to perform image-to-image translation in 3D medical images without patch cropping maintaining the original size. Additionally, our proposed model enables the generation of multiple target modalities from a single source modality. We achieve this by the proposed multiple-switchable block, noted as MS-SPADE block, which dynamically transforms source latents to target-like latents according to style. By utilizing conditioning in LDM, we can generate the desired target modality more faithfully. The pipeline of our model is illustrated in Figure 1.

Our main contributions are summarized as follows:

- We propose a model based on the LDM that translates a single source modality to various target modalities (one-to-many) in 3D medical images.
- We introduce a switchable block that transfers the source latents to target-like latents using style transfer to enhance the performance of image-to-image translation.
- We validated our method on the BraTS2021 and IXI datasets and despite our model’s ability to generate multiple modalities, it demonstrated the highest image translation performance on both datasets.

2. Related Works

Generative adversarial networks have achieved great success in image-to-image translation by performing continuous adversarial training between the generator and discriminator [7, 15, 19, 23, 46]. The generator aims to generate as realistic images as possible, while the discriminator learns to distinguish between the generated images and real images [15]. For instance, Pix2Pix [19] focuses on pixel-to-pixel image synthesis based on paired data enhancing the pixel-to-pixel similarity between real and synthesized images. CycleGAN [46] is a generalized conditional GAN enabling image synthesis even with unpaired data. NICEGAN [7] proposes a more compact architecture by sharing the encoder parts of the generator and discriminator. Compared to conventional GAN, RegGAN [23] adds a registration network to perform image-to-image translation while maintaining anatomical structures in medical images. Ea-GAN [5] is a 3D-based network designed to enhance cross-modality MR image synthesis by integrating edge information, capturing both voxel-wise intensity and image structure details. ResViT [11] is an adversarial network that incorporates a vision transformer architecture to capture contextual features in medical image translation tasks. However, these methods all work well in situations where there is a one-to-one mapping between the source and target modalities. To generate multiple target modalities from a single source modality, a separate model is needed for each target modality resulting in an increased number of models.
Conditional normalization layers are techniques used for style transfer and image synthesis tasks [13, 18, 28, 30]. Among them, adaptive instance normalization (AdaIN) [18] was initially used for style transfer tasks and has since been adopted in various computer vision tasks. Unlike other normalization techniques, conditional normalization layers involve normalizing layer activations to zero mean and unit variance, followed by denormalization through learned affine transformations from external data. Spatially adaptive normalization (SPADE) [28] has achieved successful image synthesis by applying spatially varying affine transformations to semantic maps. Unlike previous AdaIN, SPADE generates images that contain both style and semantic information allowing the creation of meaningful images using semantic maps, which is well applicable to 3D multi-modal medical imaging. Inspired by previous research, we adopt the SPADE module that transforms the style and semantic information of an image. Additionally, we propose the multi-switchable SPADE (MS-SPADE) module, which allows for flexible translation to various target modalities.

Diffusion probabilistic model has recently gained prominence in image generation delivering state-of-the-art results [8, 9, 12, 17, 31, 35]. Specifically, the denoising diffusion probabilistic model (DDPM) [17] with score matching has demonstrated superior performance in image generation. DDPM leverages a learned Markov chain to transform Gaussian noise distribution into the target distribution. Palette [35] introduces a model that utilizes DM by concatenating source and target images enabling various image-to-image translation tasks such as inpainting, uncropping, image restoration, and colorization without the need for architecture modifications. This model achieves high-fidelity results across different tasks. LDM [34] separates the training process into two stages for efficient processing. First, an autoencoder is trained to provide a perceptually equivalent lower-dimensional representation (e.g., source latents). Subsequently, DM is trained in this latent space reducing complexity and enabling efficient image generation. Importantly, this approach allows for efficient training without the need for patch cropping in 3D images.
Figure 3. (a) illustrates the training process of the autoencoder during the image compression phase to compute source latents, which involves five different losses: reconstruction, quantization, adversarial objective, perceptual, and cycle consistency. (b) shows the training process of the diffusion model, where the autoencoder is frozen, and only the UNet of the diffusion model is trained. The forward diffusion process is fixed and the input to the UNet consists of the concatenated two latents.

3. Method

To overcome the limitations of single-stage generation approaches in recent generative models, numerous studies [10, 14, 32, 33, 42] have adopted a two-stage approach to combine the strengths of different methods resulting in the development of models with superior performance. Our method trains image-to-image translation in a two-stage process as depicted in Figure 2. In the first stage, we train an autoencoder, inspired by previous research methods such as VQ-VAEs and VQGAN [14,33,39], to compress the image into a latent representation that captures its essential features. The latent space, which captures the perceptual representation of the original images, is regularized via vector quantization and allows for style transfer through conditional normalization layers [28]. We also train SPADE block to transform the style of the latents into those of the target modality (as shown in Figure 2 (a)). The proposed MS-SPADE block learns the style parameter of the target and transfers the style from the source to the target. While translation is possible using only the SPADE module, there is a noticeable difference between the true target latents and those obtained through SPADE and thus we note the output of SPADE as target-like latents. To address this discrepancy, we utilize the LDM to reduce the gap via denoising between these two latents, which is the key idea in our model. Subsequently, we combine the target-like latents (out of MS-SPADE) and noisy target latents (out of forward diffusion), similar to the palette approach, to train the DM (as shown in Figure 2 (b)). Our model prioritizes minimizing changes in the source image by first adjusting its latent style and then focusing on generating target-specific latents.

3.1. Image compression to compute latents

Similar to previous work [14, 31, 34], we employ a perceptual compression model to facilitate the application of the diffusion model in the latent space. We also introduce the MS-SPADE block in the latent space to convert the source latents to the target-like latents. The switchable normalization is illustrated in Figure 2 (a). Our SPADE block learns the mean and standard deviation parameters of the target during training. Thus, during inference, it does not require the target as input but instead uses the target’s style parameters for normalization.

The MS-SPADE block performs different normalizations according to the target modality enabling the transformation of the transformed latents to follow the desired distribution. Let \( h \in \mathbb{R}^{N \times C \times H \times W \times D} \) be the input of SPADE Block, where \( N, C, H, W \) and \( D \) are the size of batch, channel, height, width and depth respectively. The SPADE can be applied as follows:

\[
x_{n,c,h,w,d} = \gamma_{c,h,w,d}(h) \frac{h_{n,c,h,w,d} - \mu_c}{\sigma_c} + \beta_{c,h,w,d}(h)
\]

where \( x_{n,c,h,w,d} \) denotes the \( n, c, h, w, d \)-th element of the feature tensor \( h \), \( \mu_c \) and \( \sigma_c \) are the mean and standard deviation of the latent features in channel \( c \), and \( \gamma_{c,h,w,d} \) and \( \beta_{c,h,w,d} \) are modulation parameters that are learned and applied differently according to the target modality during training.

Furthermore, our compression model is based on VQ-GAN [14] and is trained using reconstruction loss, quantization loss, perceptual loss based on previous research and patch-based adversarial objectives [14, 22, 39]. Additionally, we incorporate cycle consistency loss to ensure that the original input can be reconstructed back to the original image through the cycle path [46]. The training process is illustrated in Figure 3 (a).

3.2. Diffusion Model

LDM achieves high-resolution image synthesis while reducing computational cost by training the diffusion process in the compressed latent space [34]. Additionally, Palette demonstrates high-quality image synthesis in tasks such as
The loss function is defined as follows:

$$\mathcal{L}_{diff} := \mathbb{E}_{x \sim \mathcal{N}(0,1), t} \left[ \| \epsilon - \epsilon_\theta(z_{src}, z_{tar}, t) \|^2_2 \right]$$  \hspace{1cm} (2)

Our model’s neural backbone, denoted as $\epsilon_\theta(z_{src}, z_{tar}, t)$, is implemented as a time-conditioned UNet [17, 34]. The forward diffusion process is fixed and $z_{src}$ and $z_{tar}$ can be efficiently obtained from $\mathcal{E}$ during training.

3.3. Modality Conditioning

Inspired by the successful utilization of cross-attention as a conditional image generator in previous studies [8, 34], we incorporate modality conditioning into our model to enhance the image translation to the target modality. Our approach involves converting the given modality into a one-hot vector $y$ and using it as input for the key and value components of cross-attention during training. The cross-attention in the UNet operates as follows:

$$
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^\top}{\sqrt{d}} \right) \cdot V
$$

Here, $Q_{i} \in \mathbb{R}^{d \times d_y}$ denotes a intermediate representation of the UNet implementing $\epsilon_\theta$ and $W_Q^{(i)} \in \mathbb{R}^{d \times d_y}$ and $W_V^{(i)} \in \mathbb{R}^{d \times d_y}$ are learnable projection matrices [20, 34, 41]. We redefine the loss function with modality conditioning, expressed as $\mathcal{L}_{diff}$ with $y$ representing the one-hot vector for the given modality class.

$$
\mathcal{L}_{diff} := \mathbb{E}_{x \sim \mathcal{N}(0,1), t} \left[ || \epsilon - \epsilon_\theta(z_{src}, z_{tar}, t, y) ||^2_2 \right]  \hspace{1cm} (3)
$$

For the inference process, we utilize the encoder acquired from the image compression phase. We obtain a target-similar latent by passing the source image through the encoder and the MS-SPADE block. This latent is then concatenated with random noise and used as the conditioning for the diffusion model, facilitating the image translation into the target modality.

4. Experiments

4.1. Datasets

To evaluate the effectiveness of our model, we utilized the multi-modal brain tumor segmentation challenge 2021 (BraTS 2021) dataset [1–3, 27]. We trained our model using the BraTS 2021 training dataset, which consists of 1251 colorization and inpainting reconstruction without relying on conditioning [35]. Instead, it simply takes the target image as an additional input to the DM resulting in superior results. We employ trained perceptual compression models, which consist of an encoder $E$ and a decoder $D$, to apply the diffusion model in the latent space [34]. Inspired by previous studies [34, 35], we design an LDM that takes the source latents transformed to be similar to the target latents besides noisy target latents as inputs. This process is illustrated in Figure 2 (b). The latent $z_{tar}$ represents the source latents transformed to be similar to the target latents by the SPADE block. It is denoted as target-like latents to emphasize the gap between true target latents and those from SPADE. $z_{tar}$ denotes the noisy input latents at time step $t$. Our diffusion model takes $z_{src}$ and $z_{tar}$ as inputs at time step $t$ to perform the task of predicting the noise at that time step. During the training process, the autoencoder used for image compression is frozen, and only the diffusion model is trained (as shown in Figure 3 (b)). The loss function is defined as follows:

$$
\mathcal{L} = L_{diff} + \lambda \cdot L_{adv}
$$

where $L_{diff}$ and $L_{adv}$ are the diffusion loss and the adversarial loss, respectively.

4.2. Experiments

4.2.1. Datasets

We utilized the multi-modal brain tumor segmentation challenge 2021 (BraTS 2021) dataset [1–3, 27]. We trained our model using the BraTS 2021 training dataset, which consists of 1251 colorization and inpainting reconstruction without relying on conditioning [35]. Instead, it simply takes the target image as an additional input to the DM resulting in superior results. We employ trained perceptual compression models, which consist of an encoder $E$ and a decoder $D$, to apply the diffusion model in the latent space [34]. Inspired by previous studies [34, 35], we design an LDM that takes the source latents transformed to be similar to the target latents besides noisy target latents as inputs. This process is illustrated in Figure 2 (b). The latent $z_{tar}$ represents the source latents transformed to be similar to the target latents by the SPADE block. It is denoted as target-like latents to emphasize the gap between true target latents and those from SPADE. $z_{tar}$ denotes the noisy input latents at time step $t$. Our diffusion model takes $z_{src}$ and $z_{tar}$ as inputs at time step $t$ to perform the task of predicting the noise $\epsilon$ at that time step. During the training process, the autoencoder used for image compression is frozen, and only the diffusion model is trained (as shown in Figure 3 (b)). The loss function is defined as follows:

$$
\mathcal{L} = L_{diff} + \lambda \cdot L_{adv}
$$

where $L_{diff}$ and $L_{adv}$ are the diffusion loss and the adversarial loss, respectively.
subjects and includes four MRI modalities (T1, T1ce, T2, FLAIR). For evaluating the image translation capability of our model, we employed the BraTS 2021 validation dataset, which comprises 219 subjects. We also validated our model on the IXI dataset [4], which includes T1, T2, and PD modalities. Out of 574 subjects, 459 were used for training and 115 for testing. For details on the dataset, please refer to the supplementary.

### 4.2. Implementation Details

The image compression in our model was implemented using VQGAN [14] and the diffusion model was based on LDM [34]. The implementation was done using PyTorch\(^1\) and MONAI\(^2\) libraries. For detailed information on the model architecture, refer to the supplementary materials. As for the hyperparameters, we used the AdamW [26] optimizer with an learning rate of $2 \times 10^{-6}$, we set the number of time steps $T=1000$ with scaled-linearly scheduled noise levels ranging from 0.0015 to 0.0195. The training was conducted on four A100 80GB GPUs with a batch size of 1 per GPU. Additional information on the training hyperparameters can be found in the supplementary materials.

### 4.3. Evaluation Metrics

We compared the performance of methods with previous approaches commonly used in medical image translation research [43]. The synthesis quality was evaluated using peak signal-to-noise ratio (PSNR), normalized mean squared error (NMSE), and structural similarity index (SSIM) [43]. They were computed between the ground truth images and the synthesized target images. The average and standard deviation of the metrics were reported on an independent test set that did not overlap with the training-validation sets. The evaluation was performed on 3D volumes for all tests. For the 2D methods, the synthesized target images were stacked to form a 3D volume for comparison.

### 5. Results

#### 5.1. Comparison of Ours to Baselines

Table 1 presents the quantitative evaluations of each method for three tasks: $T1 \rightarrow T2$, $T2 \rightarrow FLAIR$, and $T1 \rightarrow PD$, across two different datasets. These tasks represent real-world scenarios where each task requires one-to-one image synthesis, and our model is capable of one-to-many image synthesis and

\(^{1}\)https://pytorch.org/

\(^{2}\)https://monai.io/
Figure 6. The figures showcase the image translation results from each source modality to the corresponding target modality using our proposed model for all possible combinations. Each row corresponds to a source modality, each column represents a target modality and the last row represents the ground truth.

Table 2. The values present the quantitative evaluation of image translation results from source modalities to target modalities using our proposed model. The diagonal entries denote reconstructing from latents and are presented for completeness.

5.2. Multi Modalities Image Translation

We conducted image translation experiments using our proposed model with different scenarios of translating one source modality to another target modality given four MRI modalities. Figure 6 visualizes the results of image translation tasks from each source modality to the target modality demonstrating successful image synthesis in all cases for the BraTs dataset. The diagonal entries represent tasks where the source and target modalities are the same. The quantitative evaluation of each task in Table 2 reveals the performance of our proposed model for the BraTs dataset. When the target is T1, the best translation performance is...
Figure 7. The figures depict the results of the $T_1 \rightarrow T_2$ task for each ablation method on the BraTS2021 dataset. Each method represented in the figure corresponds to the methods listed in Table 3.

Table 3. The values present the quantitative evaluation metrics of the $T_1 \rightarrow T_2$ task for different ablated models: diffusion model, palette, MS-SPADE block, and image compression model for the BraTs dataset. The results are shown and the values in bold indicate the highest performance for each task. Reg-type is the regularization type.

<table>
<thead>
<tr>
<th>Method</th>
<th>Diff</th>
<th>Palette</th>
<th>SPADE</th>
<th>Reg-type</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>✓</td>
<td></td>
<td>VQ</td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td>✓</td>
<td>✓</td>
<td>VQ</td>
<td></td>
</tr>
<tr>
<td>(c)</td>
<td>✓</td>
<td>✓</td>
<td>VQ</td>
<td></td>
</tr>
<tr>
<td>(d)</td>
<td>✓</td>
<td>✓</td>
<td>KL</td>
<td></td>
</tr>
<tr>
<td>ours</td>
<td>✓</td>
<td>✓</td>
<td>VQ</td>
<td></td>
</tr>
</tbody>
</table>

Achieved when the source is T1ce. Similarly, for the target of T1ce, the source of F1 yields the highest performance. When the target is T2, using FLAIR as the source produces the best results. Lastly, when the target is FLAIR, the translation performance is highest when the source is T1ce. T1ce, being an MRI with contrast agent administration, contains abundant information about anatomy and tumors compared to other modalities, which explains its superior performance as a source modality in image translation. However, obtaining T1ce, which requires injecting contrast agents can be challenging in many situations. Therefore, our results demonstrate the feasibility of utilizing alternative modalities for image translation when obtaining T1ce is difficult.

5.3. Ablation Study

To demonstrate the incremental value of design components in our model, we conducted several lines of ablation studies. Specifically, we evaluated the performance contribution of the diffusion model, palette module, MS-SPADE block, and image compression model focusing on the $T_1 \rightarrow T_2$ task using the BraTS dataset. Table 3 presents the quantitative evaluation results, while Figure 7 illustrates the image synthesis results of each method in the ablation study for the BraTs dataset. In Table 3, (a) represents the performance of the translation when using the diffusion model alone demonstrating that image translation is achievable. However, (b) with the addition of the palette shows superior performance. Additionally, (c) shows the results of conducting style transfer using only the MS-SPADE block without the diffusion model indicating limitations in image translation performance. (d) and “ours” represent the results with different regularization types of the image compression model applied in previous studies [34]. KL-reg, although computationally expensive for 3D training, had an impact on performance due to the adjustment of model size and resulting in slightly blurry compression results. It was observed that the best translation performance was achieved when all the elements were used together.

6. Conclusion

Our method is an important application of computer vision in medicine. We propose a model for multi-modal image translation and conducted comprehensive experiments to evaluate its performance. Our model demonstrated successful image synthesis across different source and target modalities showcasing its versatility. By comparing our method with existing approaches, our method outperformed them in both quantitative and qualitative evaluations. Furthermore, our model showcased the ability to perform one-to-many image synthesis, going beyond the limitations of one-to-one tasks performed by other models. We also showed that our model achieved excellent performance even without pre-processing steps such as patch cropping and could successfully perform image synthesis on 3D medical images. Due to the 3D nature of the proposed approach, our method may be computationally expensive. Future work includes translating from more than two source modalities to a chosen target modality and validating on other medical imaging such as computed tomography.
References


