DAMM-Diffusion: Learning Divergence-Aware Multi-Modal Diffusion Model for Nanoparticles Distribution Prediction

Supplementary Material

6. Method

6.1. Preliminaries of Diffusion Models

Diffusion Models are consisted of two processes: the forward process and the reverse process. The forward process progressively perturbs x_0 to a latent variable by adding noise sampling from isotropic Gaussian distribution. Mathematically, a T-step forward process can be formulated as the following Markovian chain:

$$q(x_1, ..., x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{16}$$

where $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$ is a normal distribution whose mean value is $\sqrt{1-\beta_t}x_{t-1}$ and the deviation is $\beta_t\mathbf{I}$. Here, β_t is the variance schedule across diffusion steps. The latent variable $x_T \sim \mathcal{N}(0, \mathbf{I})$ when $T \to \infty$.

The reverse process can be viewed as a corresponding denoise process to recover x_0 from the latent variable x_T, which can be parameterized as:

$$p_\theta(x_0, ..., x_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \tag{17}$$

where $p_\theta(x_{t-1}|x_t)$ is represented as the approximate Gaussian such that $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$, $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t))$ are the mean and variance which can be estimated by θ. In practice, the variance is set to untrained time dependent constants i.e., $\Sigma_\theta(x_t, t) = \beta_t\mathbf{I}$.

The objective of the Diffusion Model is to maximize the Evidence Lower Bound (ELBO) of the joint distribution of forward process, which can be simplified as:

$$\mathbb{E}_{x_0,t,\epsilon}||\epsilon - \epsilon_\theta(x_t, t)||_2^2, \tag{18}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is the Gaussian noise added in x_t and θ represents the parameter of a neural network.

Conditional Diffusion Models (CDMs) aim to implement controllable diffusion with condition y for jointly training, and the objective can be modified as:

$$\mathbb{E}_{x_0,y,t,\epsilon}||\epsilon - \epsilon_\theta(x_t, y, t)||_2^2. \tag{19}$$

For the image-to-image translation task, the condition y is the image in the source domain.

Latent Diffusion Models (LDMs) [29] operate the forward and reverse processes in a latent space rather than the original pixel space which help focus on the important semantic information of the data while mitigating the need for redundant and intensive computations.

6.2. Computational Complexity

Here, we focus on analyzing the computational complexity of the MMFM and UAFM modules. MMFM is consisted of two parts i.e., spatial attention and channel attention. The computational complexity for the spatial attention with the input feature $v \in R^{C\times H\times W}$ in Eq. (6) is $2O(C^2 \times K^2 \times H \times W) + 4O(C \times H \times W)$, where the computational complexity for each convolution, normalization and activation operations are $O(C^2 \times K^2 \times H \times W)$, $O(C \times H \times W)$ and $O(C \times H \times W)$, respectively. Similarly, the computational complexity of the channel attention with the concatenated input $f_{sp} \in R^{2C\times H\times W}$ in Eq. (7) is $O(2C \times H \times W) + O(\frac{2C}{r} \times 2C) + O(2C \times \frac{2C}{r})$, where the complexity of AvgPool and linear operations are $O(2C \times H \times W)$ and $O(\frac{2C}{r} \times 2C) + O(2C \times \frac{2C}{r})$, respectively. **In summary, the total computational complexity of the MMFM module is** $2O(C^2 \times K^2 \times H \times W) + 4O(C \times H \times W) + O(2C \times H \times W) + O(\frac{2C}{r} \times 2C) + O(2C \times \frac{2C}{r})$. On the other hand, UAFM mainly involves the calculation of uncertainty-aware cross-attention (shown in Eq. (9)). Thus, **the total complexity for UAFM module is** $2O(H^2 \times W^2 \times d) + O(H^2 \times W^2)$.

7. Additional Experiments

7.1. Additional Results of Uni-modal Methods

We present additional results for each individual modality (i.e., nuclei and vessels) of uni-modal methods in Tab. 8. We can observe that our method consistently outperforms the uni-modal methods for both nuclei and vessels. The results also support the finding that vessels are more beneficial for predicting NPs.

7.2. Settings of Hyperparameters.

We conduct studies about the hyperparameters of λ in Eq. (13) and α in Eq. (14) on the internal validation, with results in Fig. 6 and Fig. 7. Based on outcomes across different datasets, we find that λ achieves optimal performance at 1e-4, while setting α to 0.1 is more beneficial for the results.

7.3. Different Types of Datasets and Tasks

Dataset. We further validate the effectiveness of DAMM-Diffusion on the brain image synthesis task. Specifically, we test our DAMM-Diffusion on the Multi-modal Brain Tumor Segmentation Challenge 2018 (BRATS) dataset [27].

Table 8. Performance comparisons with uni-modal methods on internal validation set. The symbol * indicates significant improvement ($p < 0.05$).

Methods	SSIM % (nuclei)	SSIM % (vessels)	PSNR (nuclei)	PSNR (vessels)
Cyclegan	74.87±3.64 *	84.07±2.67 *	28.12±2.46 *	36.96±2.34 *
Pix2pix	76.02±3.24 *	87.81±2.15 *	31.27±2.12 *	38.97±2.70 *
LDM	78.23±1.02 *	92.97±0.65 *	32.57±1.17 *	43.72±0.62 *
BBDM	79.05±1.31 *	93.01±0.81 *	32.34±0.96 *	43.96±0.75 *
Ours	**96.54±0.62**		**47.93±0.67**	

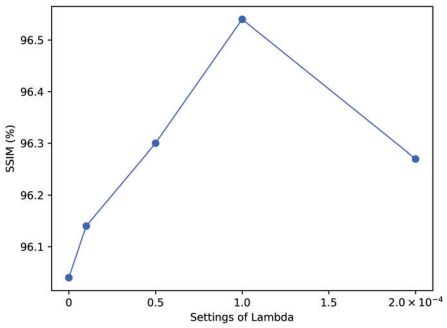

Figure 6. The effect of hyperparameter λ.

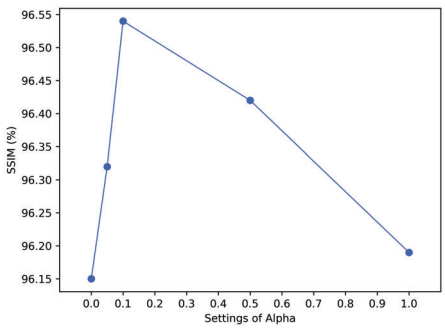

Figure 7. The effect of hyperparameter α.

The BRATS dataset consists of 285 patients with the multi-modal MRI scans including different imaging modalities: T_1, T_2 and $FLAIR$. These scans were acquired using various clinical protocols and scanners from 19 different institutions, ensuring a diverse and comprehensive dataset. Each modality volume has a size of 240 × 240 × 155 voxels. In this study, we automatically select 2D axial-plane slices, crop a central 200 × 200 region from each and then resize it to 256 × 256. Additionally, we randomly split the 285 subjects to 80% for training and 20% for testing.

Results of Different Modalities on Uni-modal Branch. We compare the performance of DAMM-Diffusion when using different input modalities in the uni-modal branch on the BRATS dataset. As shown in Tab. 9, the results indicate that the choice of different input images do not significantly impact the final performance on the BRATS dataset. This may be due to the fact that each modality in the multi-modal brain image synthesis effectively contributes to the overall outcomes.

Task	$T_1, T_2 \rightarrow$ FLAIR		T_1, FLAIR $\rightarrow T_2$		T_2, FLAIR $\rightarrow T_1$	
Input	T_1	T_2	T_1	FLAIR	T_2	FLAIR
SSIM	88.90 ±6.25	89.20 ±5.65	92.37 ±3.66	92.86 ±3.60	92.68 ±4.22	92.27 ±4.42
PSNR	24.13 ±3.38	24.23 ±3.16	25.07 ±3.74	25.25 ±3.48	25.61 ±2.49	25.29 ±2.60

Table 9. Effects of choosing different types of images in the uni-modal branch on the BRATS Dataset.

Qualitative Results. We present the representative target images for T_1, $T_2 \rightarrow$ FLAIR , T_2, FLAIR $\rightarrow T_1$ and T_1, FLAIR $\rightarrow T_2$ in Fig. 8, Fig. 9 and Fig. 10, respectively. Compared to the baseline methods, our approach generates target images with significantly reduced artifacts and enhanced clarity in tissue depiction. As shown in Fig. 8, DAMM-Diffusion can accurately capture brain lesions and provide the details of pathological regions, while the other methods fail to achieve. These results demonstrate the superiority of DAMM-Diffusion in generating the reliable medical images.

7.4. Additional Visualization Analysis

We provide more visualization results on the NPs distribution prediction task, including the generated whole-slide and patch-level images in Fig. 11 and Fig. 12, respectively.

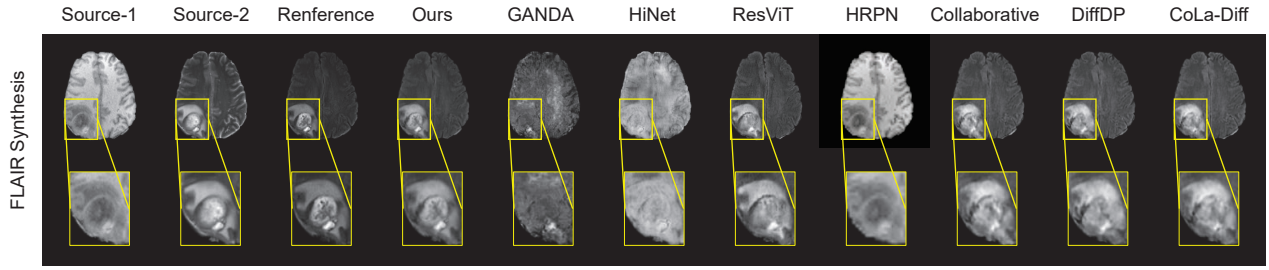

Figure 8. Visualization results of benchmark methods and DAMM-Diffusion on the BRATS dataset for the representative many-to-one synthesis task: T_1, $T_2 \rightarrow$ FLAIR.

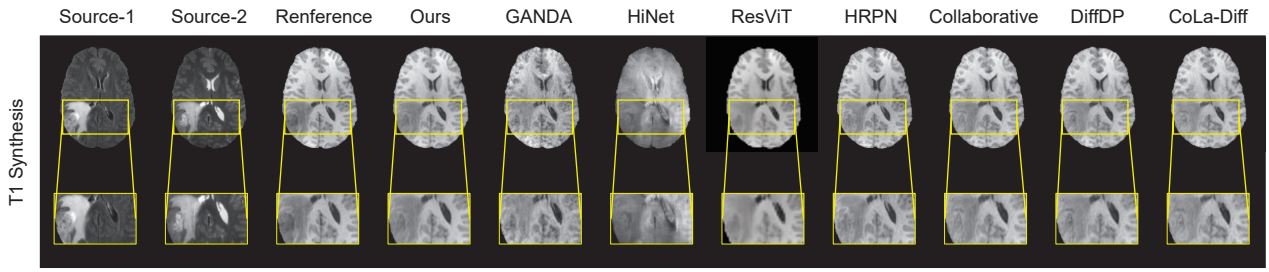

Figure 9. Visualization results of benchmark methods and DAMM-Diffusion on the BRATS dataset for the representative many-to-one synthesis task: T_2, FLAIR $\rightarrow T_1$.

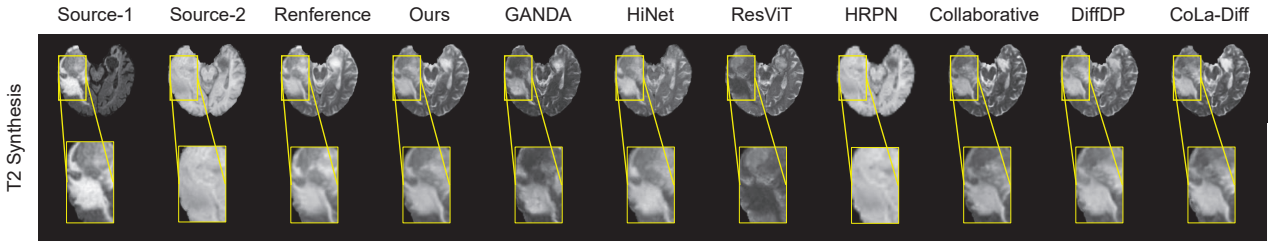

Figure 10. Visualization results of benchmark methods and DAMM-Diffusion on the BRATS dataset for the representative many-to-one synthesis task: T_1, FLAIR $\rightarrow T_2$.

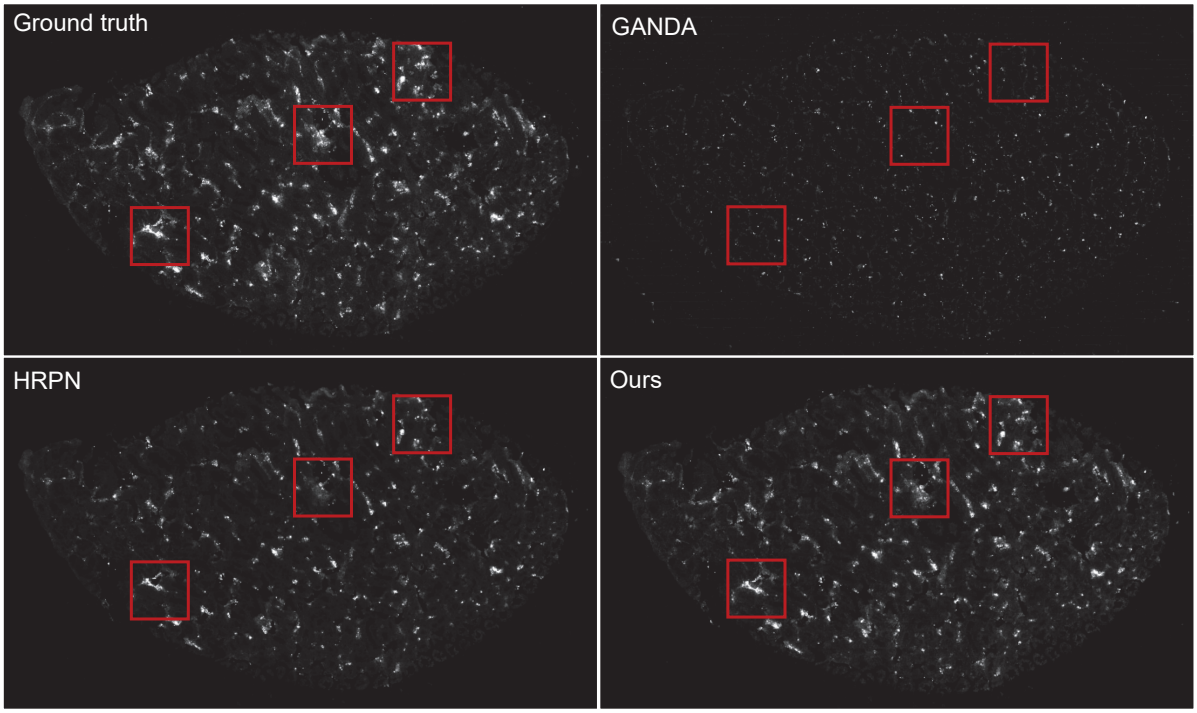

Figure 11. Qualitative comparison between the proposed method and the previous methods for NPs distribution prediction in a whole-slide image.

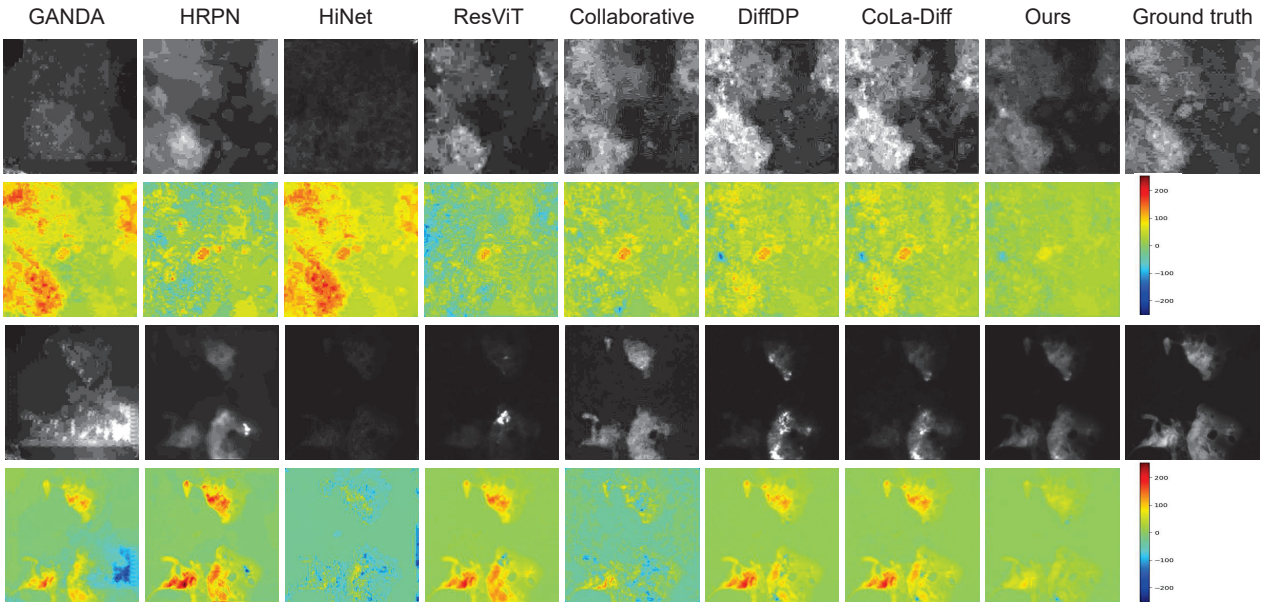

Figure 12. Visualization of generated NPs distribution (1st and 3th rows) and corresponding difference maps (2nd and 4th rows) at patch level.