CDM: A Cross-conditioned Diffusion Model for Multi-modal Medical Image Synthesis

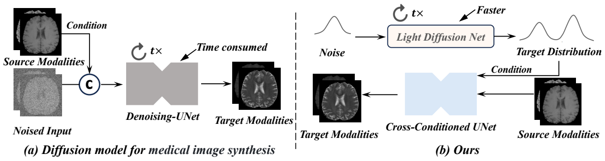
Anonymous Author
Affliation
Address
email

Abstract—**Multi-modal magnetic resonance imaging (MRI) provides rich, complementary information for analyzing diseases. However, the practical challenges of acquiring multiple MRI modalities, such as cost, scan time, and safety considerations, often result in incomplete datasets. This affects both the quality of diagnosis and the performance of deep learning models trained on such data. Recent advancements in generative adversarial networks (GANs) and denoising diffusion models have shown promise in natural and medical image-to-image translation tasks. However, the complexity of training GANs and the computational expense associated with diffusion models hinder their development and application in this task. To address these issues, we introduce a Cross-conditioned Diffusion Model (CDM) for medical image-to-image translation. The core idea of CDM is to use the distribution of target modalities as guidance to improve synthesis quality while achieving higher generation efficiency compared to conventional diffusion models. First, we propose a Modality-specific Representation Model (MRM) to model the distribution of target modalities. Then, we design a Modality-decoupled Diffusion Network (MDN) to efficiently and effectively learn the distribution from MRM. Finally, a Cross-conditioned UNet (C-UNet) with a Condition Embedding module is designed to synthesize the target modalities with the source modalities as input and the target distribution for guidance. Extensive experiments conducted on the BraTS2023 and UPenn-GBM benchmark datasets demonstrate the superiority of our method.**

Index Terms—**Diffusion model ; Multi-modal MRI ; Medical image to image translation ; Medical image Synthesis.**

I. Introduction

MULTI-MODAL magnetic resonance imaging (MRI) is crucial for the comprehensive analysis and diagnosis of diseases and is routinely used in clinical settings [27], [28], [37], [14], [16], [31], [33], [30]. They provide rich, complementary information for analyzing brain tumors. Specifically, for gliomas, the commonly used MRI sequences including T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuation Inversion Recovery (T2-FLAIR) images [32], [18]. Each sequence plays a varying role in distinguishing between the tumor, peritumoral edema, and the tumor core. However, obtaining multiple modalities in clinical settings can be challenging due to factors such as scan costs, limited scan time, and safety considerations. Consequently, the absence of certain crucial modalities can have a detrimental impact on the quality of diagnosis and treatment. Furthermore, deep learning models based on multi-modal MRIs also suffer from decreasing performance when crucial modalities are missing in training data.



(a) Diffusion model for medical image synthesis

(b) Ours

Fig. 1: Comparison between the conventional Diffusion model (a) and our method (b). Our method replaces the time-consuming denoising UNet with a light Diffusion network, which achieves higher efficiency.

Generative adversarial networks (GANs) [9], [38], [5], [15] have been extensively explored for natural image-to-image translation. However, these methods are difficult to apply directly to medical imaging due to the domain gap between natural and medical images. To address this issue, RegGAN [11] uses an additional registration network to fit the misaligned noise distribution. ResViT [3], [17] proposes a transformer-based central bottleneck module designed to distill task-critical information while preserving both global and local information within high-dimensional medical images. Although there has been significant development in these methods, the training process of GANs is not stable.

Recently, denoising diffusion models [6], [35], [25], [8], [20], [36], which are capable of offering better details, have shown significant success in various generative tasks.

Dhariwal [4] et al. propose the first diffusion model with conditional input and achieves better performance compared to GANs. However, diffusion models introduce additional computational costs since they must sample multiple times during inference. RCG [12] introduces the concept of self-conditioned image synthesis for the first time and outperforms conventional diffusion models in terms of accuracy and efficiency. However, RCG cannot be directly applied to image-to-image translation tasks due to its self-conditioning mechanism. Moreover, the encoder and decoder pre-trained on natural images in RCG are not well-suited for medical images.

In this paper, we propose a Cross-conditioned Diffusion Model for medical image-to-image translation, named CDM. Instead of directly sampling the target modalities in the image domain like the conventional diffusion model, CDM first samples the distribution of target modalities in latent variable space, and then this distribution is used as a condition to generate the target modalities in the image domain. First, we design

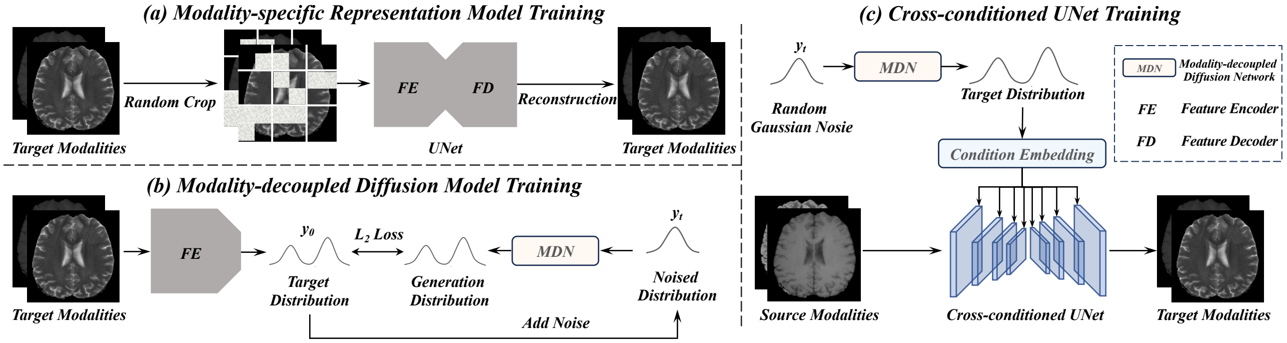


Fig. 2: An overview of the proposed Cross-conditioned Diffusion Model (CDM). First, we introduce the Modality-specific Representation Model (a) to learn the distribution of target modalities. Then, the Modality-decoupled Diffusion Network (b) is employed to learn the target distribution. Finally, the Cross-conditioned UNet (c) incorporates the source modalities and samples the target distribution as guidance to generate the target modalities.

a Modality-specific Representation Model (MRM) to learn the distribution of target modalities. Subsequently, as illustrated in Fig. 1, we replace the time-consuming denoising UNet with a light Diffusion network, called Modality-decoupled Diffusion Network (MDN), which achieves higher efficiency in both training and inference and model the target distribution from MRM. Finally, we propose a Cross-conditioned UNet (C-UNet) with a Condition Embedding module to receive the source modalities and distribution sampled by MDN as input to generate the target modalities. Extensive experiments on BraTS2023 [19], [1], [10] and UPenn-GBM [2], [13] datasets demonstrate the superiority of our proposed CDM.

II. Method

Our CDM primarily consists of three components: 1) the modality-specific representation model, which learns the distribution of target modalities; 2) the modality-decoupled diffusion network, designed for improved feature representation and efficiency; and 3) the cross-conditioned UNet model, which generates target modalities from source modalities and sampled target distribution.

A. Representation Learning for Target Modalities

Modality-specific Representation Model (MRM) Training In RCG, both the feature encoder and decoder are pre-trained on natural images, which results in reduced performance when processing medical images. We design a modality-specific representation model consisting of a feature encoder FE and decoder FD. Similar to SimMIM [29], as shown in Fig. 2 (a), we randomly mask some patches in each target modality separately and concatenate them at the channel dimension as input. Then, the MRM learns to restore the original target modalities, supervised by the L_2 loss function.

$$\mathcal{L}_{\text{MRM}} = \frac{1}{|R|} \sum_{r \in R} ||p_r - \hat{p}_r||_2, \tag{1}$$

where R denotes masked patches in target modalities, $|R|$ denotes the number of masked patches, p_r and $\hat{p_r}$ represent the prediction values and input values.

Modality-decoupled Diffusion Network (MDN) Training The core of cross-conditioned image generation lies in using the target distribution sampled by the diffusion model to guide the pixel generation process for target modalities. To achieve this, we adopt a light modality-decoupled diffusion network to efficiently sample the target distribution. As shown in Fig. 3 (a), the MDN first employs two separate linear layers to decouple the noised target distribution y_t. Then, multiple residual blocks are utilized to eliminate the noise and generate target distribution y_0. Each residual block consists of an input layer, a timestep embedding layer, and an output layer, where each layer comprises a LayerNorm [34], a SiLU [22], and a linear layer. The MDN follows Denoising Diffusion Implicit Models (DDIM) [24] for training and inference. As shown in Fig. 2 (b), during training, the target distribution y_0 from feature encoder FE is mixed with Gaussian noise ϵ over $t \in \{0, 1, ..., T\}$ steps.

$$q\left(y_t \mid y_0\right) = \mathcal{N}\left(y_t; \sqrt{\bar{\alpha}_t} y_0, \left(1 - \bar{\alpha}_t\right) \epsilon\right), \tag{2}$$

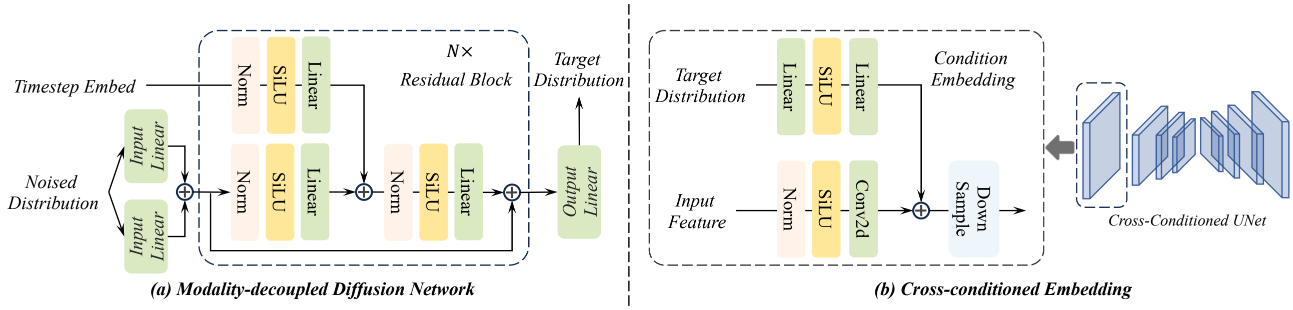
where $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s = \prod_{s=0}^{t} \left(1 - \beta_s\right)$ and β_s represent the noise schedule [6]. Then, the MDN learns to restore y_0 from noised y_t supervised by L_2 loss function. During inference, as shown in Fig. 2 (c), the target distribution y_0 is predicted by the MDN from a normal Gaussian noise y_t, along with the sample schedule [24].

$$p_\theta\left(y_{0:T}\right) = p\left(y_T\right) \prod_{t=1}^{T} p_\theta\left(y_{t-1} \mid y_t\right), \tag{3}$$

where θ denotes the training parameters of the MDN.

B. Cross-conditioned UNet (C-UNet)

To incorporate the generated target distribution into the UNet model, we propose a cross-conditioned UNet, featuring a cross-conditioned embedding module designed to merge the target distribution with the input feature at each scale. As depicted in Fig. 3 (b), for a layer in C-UNet, the input feature from source modalities is fed into a convolution block, which consists of a LayerNorm, a SiLU, and a 1×1 convolution layer, to obtain the feature representation. Simultaneously, the



(a) Modality-decoupled Diffusion Network

(b) Cross-conditioned Embedding

Fig. 3: An overview of Modality-decoupled Diffusion Network (a) and Cross-conditioned Emebedding (b).

generated target distribution is processed through an multi-layer perceptron (MLP) [26] layer, consisting of two linear layers and a SiLU, to produce the distribution representation. The feature representation is then combined with the distribution representation for fusion, and a down-sampling layer is utilized to reduce the spatial dimensions of the fused feature.

The final synthesis loss $\mathcal{L}_{\text{Syn}}$ consisting of mean square error is calculated on the prediction $\hat{Y}$ by the C-UNet and corresponding ground truth Y:

$$\mathcal{L}_{\text{Syn}} = ||\hat{Y} - Y||_2. \tag{4}$$

III. Experiments

A. Datasets and Implementation

BraTS2023 dataset The BraTS2023 dataset [19], [1], [10] contains a total of 1,251 3D brain MRI volumes. For each patient, multi-parametric magnetic resonance imaging (mpMRI) scans are available, including the four structural MRI scans: native T1-weighted (T1), post-contrast T1 (T1-Gd), native T2-weighted (T2), and T2 fluid attenuated inversion recovery (T2-FLAIR) scans.

UPenn-GBM dataset The UPenn-GBM dataset [2] is composed of 630 patients diagnosed with Glioblastoma Multiforme (GBM). Each volume also includes four modalities (namely T1, T1Gd, T2, and T2-FLAIR).

For both mentioned datasets, we denote T1Gd as T1c, and T2-FLAIR as T2f. We partition the 3D image along the Z-axis into 2D images. The T1 and T2 modalities are utilized to generate the T1c and T2f modalities.

Implementation details Our model is implemented in PyTorch 2.1.0-cuda12.1. During training, we resize each image to 256 × 256 and use a batch size of 12 per GPU for each dataset. We employ cross-entropy loss and adopt the Adam optimizer with a learning rate of 1e-4 and a decay rate of 1e-5. We run 100 epochs for all datasets. All experiments are conducted on a cloud computing platform with NVIDIA A100 GPUs. For each dataset, we randomly allocate 70% of the 3D volumes for training, and the remaining 30% for testing.

B. Comparison with SOTA Methods

We compare our CDM with seven state-of-the-art synthesis methods, including five CNN-based methods (Pix2pix [9], CycleGAN [38], GcGAN [5], CUT [21], RegGAN [11]), one transformer-based method (ResViT [3]), and one diffusion-based method (conditional diffusion model [4], denoted as 'Diffusion' in all tables). For a fair comparison, we utilize public implementations of these methods to retrain their networks, generating their best synthesis results. The Peak Signal-to-Noise Ratio (PSNR) [7], Structural Similarity Index (SSIM) [23], and Mean Absolute Error (MAE) are adopted for quantitative comparison on the BraTS2023 and UPenn-GBM datasets.

BraTS2023 Table I presents the PSNR, SSIM, and MAE scores for two modalities (T1c and T2f), along with the averaged scores of all methods on BraTS2023. Our CDM achieves the highest PSNR and SSIM scores for T1c and T2f, the lowest MAE score for T1c, and ranks second in MAE for T2f. More importantly, our method demonstrates superior quantitative performance, averaging 31.92 on PSNR, 0.941 on SSIM, and 0.0117 on MAE, respectively. Furthermore, we conduct experiments comparing the latest Diffusion method with our method, while our method outperforms the Diffusion method across all metrics.

UPenn-GBM In Table II, we list PSNR, SSIM, and MAE scores of our network and compared methods for T1c and T2f modalities on the UPenn-GBM dataset, as well as the average metrics. Among all the comparison methods, the Diffusion has the highest average PSNR and SSIM scores of 32.13 and 0.953, as well as the lowest MAE score of 0.0135. This performance is attributed to the strong representational capabilities of diffusion model. In comparison, our method has a 3.79%, 0.63%, and 13.33% improvement on PSNR, SSIM, and MAE scores, respectively, achieving state-of-the-art performance.

Visual Comparisons Fig. 4 visually compares the synthesis results predicted by our network and state-of-the-art methods on the BraTS2023 and UPenn-GBM datasets. From these visualization results, we can find that our method can more accurately synthesize the brain tumor regions than other methods and maintain the most consistent style with the ground truth. The reason behind is that our method is capable of learning modality-related information under the guidance of the distributions of T1c and T2f modalities.

C. Ablation Study

The Effectiveness of Each Module Table III lists the methods with different modules along with the average PSNR,

TABLE I: Quantitative comparison on BraTS2023 dataset

Methods	T1c			T2f			Avg		
	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓
Pix2pix [9]	27.05	0.858	0.0180	24.82	0.846	0.0250	25.93	0.852	0.0215
CycleGAN [38]	30.13	0.906	0.0120	26.85	0.883	0.0188	28.49	0.894	0.0154
GcGAN [5]	29.98	0.917	0.0129	25.98	0.872	0.0225	27.98	0.894	0.0177
CUT [21]	26.27	0.846	0.0181	23.54	0.819	0.0278	24.90	0.832	0.0229
RegGAN [11]	31.36	0.930	0.0109	29.13	0.917	**0.0135**	30.24	0.923	0.0122
ResViT [3]	31.46	0.932	0.0131	28.63	0.909	0.0166	30.04	0.920	0.0148
Diffusion [4]	31.98	0.930	0.0109	29.22	0.921	0.0155	30.60	0.925	0.0132
Ours	**33.08**	**0.948**	**0.0098**	**30.76**	**0.934**	0.0136	**31.92**	**0.941**	**0.0117**

TABLE II: Quantitative comparison on UPenn-GBM dataset

Methods	T1c			T2f			Avg		
	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓
Pix2pix [9]	28.93	0.925	0.0145	29.81	0.903	0.0206	29.37	0.914	0.0175
CycleGAN [38]	30.85	0.951	0.0122	30.25	0.914	0.0210	30.55	0.932	0.0166
GcGAN [5]	30.75	0.952	0.0133	30.41	0.928	0.0209	30.58	0.940	0.0171
CUT [21]	30.74	0.950	0.0123	31.02	0.924	0.0187	30.88	0.937	0.0155
RegGAN [11]	27.71	0.916	0.0149	27.71	0.930	0.0163	27.71	0.923	0.0156
Resvit [3]	27.43	0.930	0.0138	24.58	0.905	0.0202	26.00	0.917	0.0170
Diffusion [4]	31.84	0.962	0.0112	32.42	0.944	0.0158	32.13	0.953	0.0135
Ours	**33.04**	**0.967**	**0.0097**	**33.65**	**0.952**	**0.0138**	**33.34**	**0.959**	**0.0117**

TABLE III: Ablation study for different modules on BraTS2023 dataset.

Methods	Condition	Decouple	PSNR ↑	SSIM ↑	MAE ↓
RCG	✓		15.40	0.057	0.141
M1			29.41	0.907	0.0138
M2	✓		31.73	0.935	0.0127
Ours	✓	✓	**31.92**	**0.941**	**0.0117**

SSIM, and MAE on T1c and T2f modalities. As indicated in Table III, RCG has the lowest PSNR and SSIM scores, alongside the highest MAE score. This method employs the encoder pre-trained on natural images, which is not well-suited for medical imaging. M1 represents our basic method, which only contains a conventional UNet model. In comparison to M1, M2 integrates the distribution of target modalities (T1c and T2f) as a condition to guide the synthesis process, achieving an improvement of 7.88%, 3.08%, and 7.97% on PSNR, SSIM, and MAE, respectively. Finally, our method combines both conditional input and the MDN, achieving the state-of-the-art performance of 31.92, 0.941, and 0.0117 across the three metrics.

The Optimal Sampling Number To determine the best sampling number $N_{sampling}$, we conduct an experiment in which we increase the sampling number from 10 to 40 with a stride of 10, and evaluate the average PSNR, SSIM, and MAE for T1c and T2f. As shown in Table IV, when the sampling number reaches 40, the SSIM score decreases slightly from 0.941 to 0.940, and the improvements of PSNR and MAE are minimal. Considering the increase in computational cost from $N_{sampling} = 30$ to $N_{sampling} = 40$, we select $N_{sampling} = 30$ as our default setting.

TABLE IV: Ablation study for different sampling number $N_{sampling}$ on BraTS2023 dataset.

$N_{sampling}$	10	20	30	40
PSNR ↑	31.24	31.71	31.92	**31.93**
SSIM ↑	0.926	0.933	**0.941**	0.940
MAE ↓	0.0123	0.0123	0.0117	**0.0116**

The High Efficiency of Our CDM Fig. 5 (a) displays the frames per second (FPS) and PSNR score for both Diffusion and our CDM across different sample number N_{sample} on the BraTS2023 dataset. The larger the bubble size is, the higher the average PSNR for T1c and T2f. It is observed that our CDM surpasses the Diffusion in terms of both FPS and PSNR at different N_{sample}, indicating superior efficiency and synthesis quality.

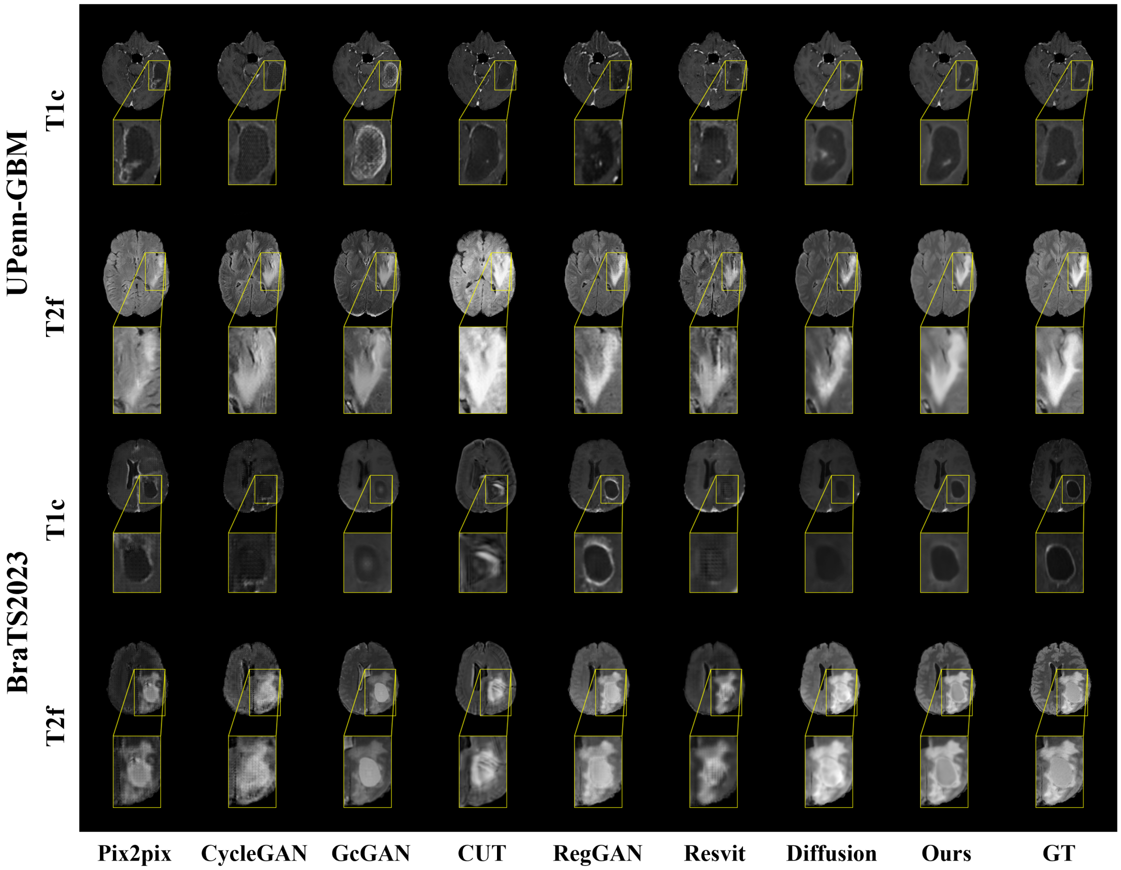
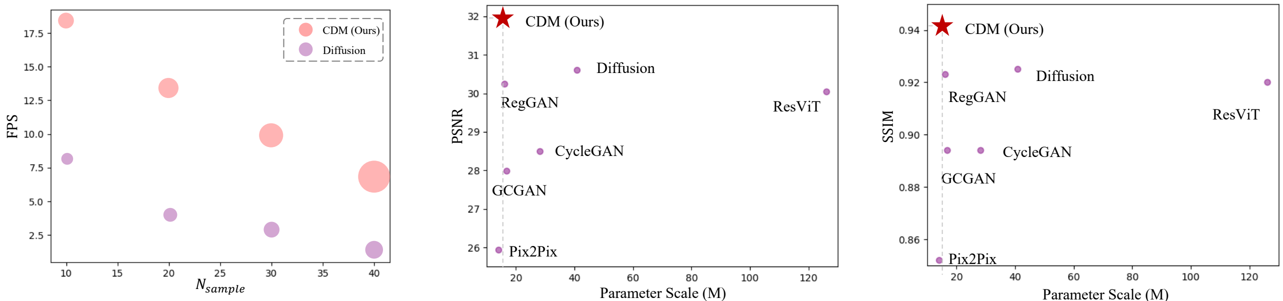


Fig. 4: Visual comparisons of proposed CDM and other state-of-the-art methods.



(a) Comparison with Diffusion (FPS & sample number N_{sample}). The bubble size indicates PSNR.

(b) Comparison with SOTA (PSNR & Parameter scale).

(c) Comparison with SOTA (SSIM & Parameter scale).

Fig. 5: The ablation studies for efficiency and parameter scale.

Comparison on Parameter Scale We compare our CDM with state-of-the-art methods in terms of parameter scale, PSNR, and SSIM on the BraTS2023 dataset. As shown in Fig. 5 (b) and (c), our method achieves new state-of-the-art results while maintaining a smaller parameter scale.

IV. Conclusion

In this paper, we propose a novel paradigm for medical image-to-image translation, named CDM. The main idea of CDM is to use a modality-specific representation model (MRM) to learn the distribution of target modalities and a modality-decoupled Diffusion network (MDN) to model the distribution from MRM while achieving higher efficiency. Finally, we propose a cross-conditioned UNet (C-UNet) to receive the source modalities as input and the distribution sampled by MDN as guidance to generate the target modalities. Extensive experiments on BraTS2023 [19], [1], [10] and UPenn-GBM [2] datasets demonstrate the superiority of our proposed CDM. The ablation studies are conducted to verify the effectiveness of each module and to demonstrate the advantages of our method in terms of efficiency and parameter scale.

References

[1] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)

[2] Bakas, S., Sako, C., Akbari, H., Bilello, M., Sotiras, A., Shukla, G., Rudie, J.D., Santamaría, N.F., Kazerooni, A.F., Pati, S., et al.: The university of pennsylvania glioblastoma (upenn-gbm) cohort: Advanced mri, clinical, genomics, & radiomics. Scientific data **9**(1), 453 (2022)

[3] Dalmaz, O., Yurt, M., Çukur, T.: Resvit: Residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging **41**(10), 2598–2614 (2022)

[4] Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)

[5] Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., Tao, D.: Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2427–2436 (2019)

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

[7] Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)

[8] Hu, M., Yan, S., Xia, P., Tang, F., Li, W., Duan, P., Zhang, L., Ge, Z.: Diffusion model driven test-time image adaptation for robust skin lesion classification. arXiv preprint arXiv:2405.11289 (2024)

[9] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)

[10] Kazerooni, A.F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., Anderson, H., et al.: The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). ArXiv (2023)

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

[12] Li, T., Katabi, D., He, K.: Self-conditioned image generation via generating representations. arXiv preprint arXiv:2312.03701 (2023)

[13] Li, W., Xiong, X., Xia, P., Ju, L., Ge, Z.: Tp-drseg: Improving diabetic retinopathy lesion segmentation with explicit text-prompts assisted sam. arXiv preprint arXiv:2406.15764 (2024)

[14] Liu, L., Aviles-Rivero, A.I., Schönlieb, C.B.: Contrastive registration for unsupervised medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems (2023)

[15] Liu, L., Hu, X., Zhu, L., Fu, C.W., Qin, J., Heng, P.A.: ψ-net: Stacking densely convolutional lstms for sub-cortical brain structure segmentation. IEEE transactions on medical imaging **39**(9), 2806–2817 (2020)

[16] Liu, L., Hu, X., Zhu, L., Heng, P.A.: Probabilistic multilayer regularization network for unsupervised 3d brain image registration. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 346–354. Springer (2019)

[17] Liu, L., Prost, J., Zhu, L., Papadakis, N., Liò, P., Schönlieb, C.B., Aviles-Rivero, A.I.: Scotch and soda: A transformer video shadow detection framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10449–10458 (2023)

[18] Luo, X., Fu, J., Zhong, Y., Liu, S., Han, B., Astaraki, M., Bendazzoli, S., Toma-Dasu, I., Ye, Y., Chen, Z., et al.: Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. arXiv preprint arXiv:2312.09576 (2023)

[19] Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)

[20] Nan, Y., Xing, X., Wang, S., Tang, Z., Felder, F.N., Zhang, S., Ledda, R.E., Ding, X., Yu, R., Liu, W., et al.: Hunting imaging biomarkers in pulmonary fibrosis: Benchmarks of the aiib23 challenge. arXiv preprint arXiv:2312.13752 (2023)

[21] Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 319–345. Springer (2020)

[22] Paul, A., Bandyopadhyay, R., Yoon, J.H., Geem, Z.W., Sarkar, R.: Sinlu: Sinu-sigmoidal linear unit. Mathematics **10**(3), 337 (2022)

[23] Sara, U., Akter, M., Uddin, M.S.: Image quality assessment through fsim, ssim, mse and psnr—a comparative study. Journal of Computer and Communications **7**(3), 8–18 (2019)

[24] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[25] Song, Y., Ermon, S.: Improved techniques for training score-based generative models. Advances in neural information processing systems **33**, 12438–12448 (2020)

[26] Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. Advances in neural information processing systems **34**, 24261–24272 (2021)

[27] Wang, H., Luo, X., Chen, W., Tang, Q., Xin, M., Wang, Q., Zhu, L.: Advancing uwf-slo vessel segmentation with source-free active domain adaptation and a novel multi-center dataset. arXiv preprint arXiv:2406.13645 (2024)

[28] Wang, H., Zhu, L., Yang, G., Guo, Y., Zhang, S., Xu, B., Jin, Y.: Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. arXiv preprint arXiv:2308.09475 (2023)

[29] Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9653–9663 (2022)

[30] Xing, Z., Wan, L., Fu, H., Yang, G., Zhu, L.: Diff-unet: A diffusion embedded network for volumetric segmentation. arXiv preprint arXiv:2303.10326 (2023)

[31] Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv preprint arXiv:2401.13560 (2024)

[32] Xing, Z., Yu, L., Wan, L., Han, T., Zhu, L.: Nestedformer: Nested modality-aware transformer for brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 140–150. Springer (2022)

[33] Xing, Z., Zhu, L., Yu, L., Xing, Z., Wan, L.: Hybrid masked image modeling for 3d medical image segmentation. IEEE Journal of Biomedical and Health Informatics (2024)

[34] Xu, J., Sun, X., Zhang, Z., Zhao, G., Lin, J.: Understanding and improving layer normalization. Advances in Neural Information Processing Systems **32** (2019)

[35] Yang, Y., Xing, Z., Zhu, L.: Vivim: a video vision mamba for medical video object segmentation. arXiv preprint arXiv:2401.14168 (2024)

[36] Ye, T., Chen, S., Chai, W., Xing, Z., Qin, J., Lin, G., Zhu, L.: Learning diffusion texture priors for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2524–2534 (2024)

[37] Zhao, J., Xing, Z., Chen, Z., Wan, L., Han, T., Fu, H., Zhu, L.: Uncertainty-aware multi-dimensional mutual learning for brain and brain tumor segmentation. IEEE Journal of Biomedical and Health Informatics (2023)

[38] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)