Proto-Gaussian: MRI Modality Translation Based on Learnable Structural Prototypes and 2D Gaussian Splatting

Zizheng Li[1], RenDong Xie[2], Huadeng Wang[1], Zhifen He[2], Bin Liu[2], Bo Li[3,2*], and Xiaonan Luo[1*]

[1] Guilin University of Electronic Technology, Guilin 541004, China
23031202017@mails.guet.edu.cn, whd@guet.edu.cn, luoxn@guet.edu.cn
[2] Nanchang Hangkong University, Nanchang 330000, China
2307070100001@stu.nchu.edu.cn, zfhe323@nchu.edu.cn, nyliubin@nchu.edu.cn
[3] Shandong Technology and Business University, Yantai 264005, China
libo@nchu.edu.cn

Abstract. Medical image modality translation is hindered by the entanglement of geometry and modality-specific textures, often causing structural distortions. Existing methods operate at the pixel or feature level, failing to separate geometry from appearance and yielding anatomically inconsistent results. We present Proto-Gaussian, a novel MRI modality translation framework that explicitly disentangles structural geometry from modality-specific appearance using 2D Gaussian image representation, surpassing conventional pixel or patch-based approaches in preserving anatomical consistency. The framework operates in two stages: first, a shared Gaussian Bank captures modality-invariant geometric prototypes through 2D Gaussian primitives and self-supervised learning; second, a Label-Adaptive Token Refinement module maps texture features across modalities while leveraging the learned geometric structures, achieving clear structure-texture separation. Finally, a fully differentiable 2D Gaussian splatting renderer synthesizes high-fidelity target-modality images, maintaining both structural accuracy and visual realism. Extensive experiments on brain MRI datasets demonstrate the superiority of our method through quantitative and qualitative comparisons with state-of-the-art approaches.

Keywords: 2D Gaussian · Medical image modality translation · Differential rendering

1 Introduction

The field of computer vision has witnessed remarkable progress in recent years, particularly in medical image analysis and generation [15, 46]. As an important branch of this domain, medical image modality translation plays a crucial role in

* Corresponding authors.

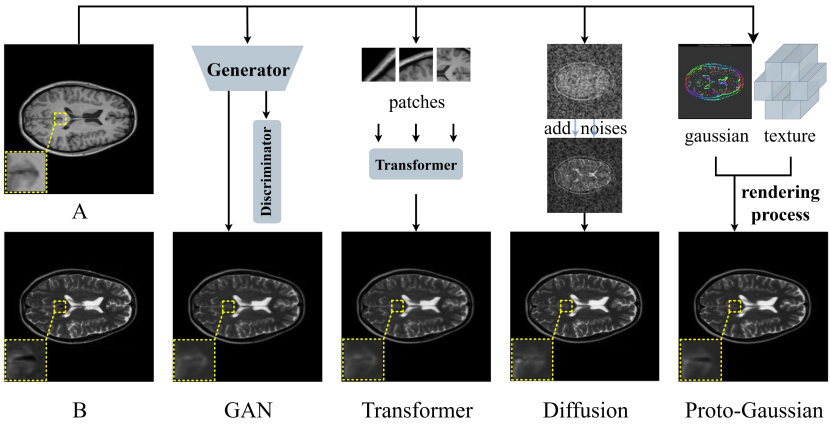

Fig. 1: Comparison of medical image modality translation methods. Existing approaches suffer from structural distortions due to implicit representations, while our method leverages explicit 2D Gaussians to disentangle structure from texture, faithfully preserving anatomical consistency.

clinical diagnosis and treatment, enabling the synthesis of missing target modalities from existing imaging data while effectively reducing the need for repeated scans and optimizing healthcare resource allocation [12,50,53]. This technology is especially significant in neurology and oncology, where complementary information from different MRI sequences (e.g., T1, T2, PD) can substantially improve diagnostic accuracy [7,55]. However, the task faces a fundamental challenge: the deep entanglement between geometric structures and modality-specific textures, which often leads to structural distortions in generated images and severely limits their clinical utility [1,52].

As illustrated in Figure 1, existing approaches for medical image modality translation have evolved along several technical pathways. Generative adversarial network-based methods learn cross-modal mappings through adversarial training [22, 33, 62], yet they suffer from inherent limitations, including training instability and mode collapse, which hinder accurate reconstruction of fine anatomical structures [23, 60]. Transformer-based hybrid architectures leverage self-attention mechanisms to enhance global dependency modeling [14, 29] and demonstrate advantages in preserving contextual information [8, 11]; however, their implicit representation schemes still fail to achieve explicit separation between structural and texture components [48, 59]. Recently proposed diffusion bridge models improve upon traditional diffusion methods [13, 18] by establishing direct inter-modality translation pathways, but exhibit limited robustness to common clinical noise and artifacts, while the high variance in intermediate steps may compromise effective information transfer [10, 35].

The common limitation of these methods lies in their operation at pixel or feature levels, failing to effectively separate geometric structures from appearance features, thereby leading to anatomically inconsistent results. Anatomical struc-

tures, as modality-invariant information shared across modalities, remain tightly coupled with modality-specific appearance features in existing implicit representation approaches, hindering genuine structure-texture decoupling [43, 44].

To address this core challenge, we propose Proto-Gaussian, a novel MRI modality translation framework based on explicit 2D Gaussian representations. Unlike traditional pixel-level or patch-level approaches, our method leverages 2D Gaussian image representations to explicitly decouple structural geometry and modality-specific appearance, demonstrating significant advantages in maintaining anatomical consistency [25, 64]. The framework operates through a two-stage pipeline: the first stage captures modality-invariant geometric prototypes through shared Gaussian repositories and self-supervised learning [27, 65]; the second stage employs a label-adaptive token optimization module to achieve precise cross-modal texture feature mapping while preserving the learned geometric structures. Finally, a fully differentiable 2D Gaussian splatting renderer synthesizes high-fidelity target modality images, ensuring both structural accuracy and visual realism [37, 63].

The main contributions of this work are summarized as follows:

- We propose the first explicit 2D Gaussian representation-based framework for medical image modality translation, achieving natural structure-texture decoupling through parametric modeling.
- We design a two-stage learning architecture that separately handles geometric prototype learning and texture feature mapping, ensuring accurate preservation of anatomical structures.
- We introduce a fully differentiable 2D Gaussian splatting renderer, providing a new technical pathway for medical image synthesis. Extensive experiments on multiple brain MRI datasets demonstrate that our method significantly outperforms state-of-the-art approaches in both qualitative and quantitative evaluations.

2 Related Work

Evolution of Generative Methods for Medical Image Modality Translation Medical image modality translation faces a fundamental trade-off between high-dimensional data modeling and anatomical structure preservation [24, 39]. Existing technical approaches have primarily developed along three directions: Generative adversarial networks have achieved breakthrough progress, where Pix2Pix [22] established the foundation for paired image translation using conditional GANs, while SAGAN [57] addressed unpaired translation through cycle-consistency loss [12, 53, 62]. However, these methods remain limited by training instability and mode collapse issues, exhibiting insufficient performance in reconstructing fine anatomical structures [1, 20, 52, 55]. Diffusion model methods have set new technical standards in generation quality and density estimation [10,18,26,35,38,43,44], where SelfRDB [2] achieved certain breakthroughs in structure preservation through diffusion bridge formulation. However, the computational burden of pixel-space operations continues to limit their application

potential in real-time clinical scenarios. Transformer architectures have brought new vitality to medical image analysis [14, 17, 29], with hybrid architectures like TransUNet [8] and ResViT [11] effectively integrating the advantages of global context modeling and local feature extraction [5, 61]. Meanwhile, multi-modal pre-training approaches such as BrainMVP [40] have explored cross-modal representation learning, but their implicit representation mechanisms cannot achieve explicit separation of anatomical structures and texture features, constituting a fundamental limitation in medical image translation tasks requiring precise structure preservation.

Multi-domain Advantages and Innovative Applications of Gaussian Representations As an explicit modeling paradigm, Gaussian representations demonstrate unique technical advantages across multiple visual tasks, providing new perspectives for addressing core challenges in medical image modality translation. In computer graphics, Gaussian splatting techniques achieve real-time efficient rendering pipelines through parametric Gaussian primitives [5, 37, 64, 65], while providing precise geometric control capabilities, significantly surpassing the efficiency limits of traditional voxel and implicit representations. In 3D reconstruction, 3D Gaussian representations substantially improve computational efficiency while maintaining reconstruction quality [3, 25, 30–32, 47, 51, 54], opening new technical pathways for rapid modeling of complex anatomical structures [6,45]. In image generation, Gaussian-based explicit representations achieve unprecedented content control precision and editing flexibility by decoupling geometric and appearance components [16, 34, 41]. These cross-domain studies collectively reveal the core advantages of Gaussian representations: parametric geometric control, inherent structure-texture separation characteristics, and efficient differentiable rendering capabilities. Based on these advantages, we innovatively introduce 2D Gaussian representations to medical image modality translation tasks, to explicitly decouple geometric structures and modality-specific textures, ensuring anatomical accuracy while achieving efficient image synthesis, providing a novel solution to overcome existing technical bottlenecks.

3 Method

A fundamental challenge in medical image modality translation lies in the deep entanglement of anatomical geometry and modality-specific textures, which makes it difficult for existing pixel- or feature-level manipulation methods to maintain anatomical structural consistency. To overcome this limitation, we propose the Proto-Gaussian framework that achieves natural decoupling of structure and texture through explicit 2D Gaussian representations. As illustrated in Figure 2, our approach employs a carefully designed two-stage architecture: Geometric Prototype Learning focuses on extracting modality-invariant structural information. At the same time, Texture-Aware Synthesis achieves high-quality modality conversion while preserving geometric consistency.

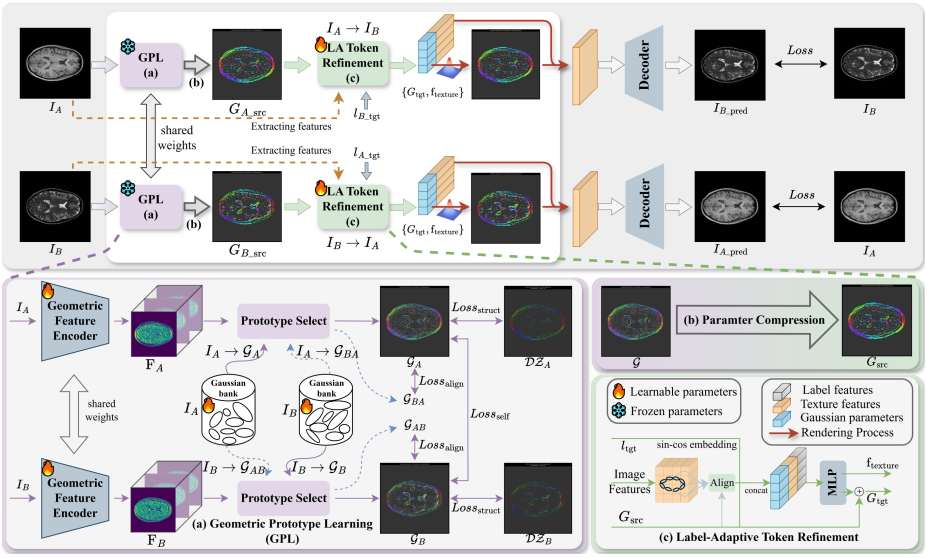

Fig. 2: The Proto-Gaussian Framework. Our framework is built around a two-stage pipeline: (a) Geometric Prototype Learning for extracting modality-invariant anatomical structures, (b) Parameter Compression that condenses the representation, and (c) Texture-Aware Synthesis that generates the final image with high-fidelity texture while strictly preserving geometric consistency.

3.1 Geometric Prototype Learning

Existing medical image modality translation methods typically perform implicit modeling in pixel space or feature space, failing to separate geometric structure from appearance characteristics effectively. This limitation often leads to anatomical structure distortion in generated images, severely impacting their clinical diagnostic utility. Inspired by differential geometry theory, we conceptualize medical images as scalar fields on 2D manifolds, where anatomical structure information is encoded in the intrinsic geometric properties of the manifold. Based on this insight, we propose a geometric prototype learning mechanism using parameterized Gaussian representations.

Gaussian Representation We introduce a novel Gaussian mixture manifold representation that establishes a mapping from image space to the Gaussian manifold through parameterized 2D Gaussian distributions:

$$\mathcal{G}(p) = (\mu(p), \Sigma(p)) \tag{1}$$

Here, $\mathcal{G}(p)$ denotes the Gaussian distribution representation at position p, $\mu(p) = (x, y)$ represents the location parameters (mean) of the Gaussian distribution encoding spatial coordinates of key anatomical landmarks, and $\Sigma(p) = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$ denotes the covariance matrix describing scale, orientation, and

shape characteristics of local geometric structures, where σ_x, σ_y indicate standard deviations along x and y directions, and ρ represents the correlation coefficient of the Gaussian distribution. This parameterized representation enables explicit separation of spatial structural information from appearance features.

Shared Geometric Prototype Bank To achieve cross-modal geometric consistency, we design a bank mechanism based on shared geometric prototypes [19]. Specifically, we construct a unified geometric prototype bank $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K\}$ that is shared across different modalities, while ensuring geometric representation consistency through carefully designed alignment constraints.

Each geometric prototype $\mathbf{p}_k = (\sigma_x^k, \sigma_y^k, \rho^k)$ represents a fundamental geometric structure pattern. We adopt a decoupled design strategy: position parameters are fixed as absolute grid coordinates $\mu_{\text{grid}}(i,j) = (x,y)$, while shape parameters are dynamically generated through the shared prototype bank. The complete geometric representation is synthesized through:

$$\mathbf{g}_{ij} = \left[\mu_{\text{grid}}(i,j),\ \sum_{k=1}^{K} \frac{\exp(\langle \phi(\mathbf{F}_{ij}), \phi(\mathbf{e}_k)\rangle/\tau)}{\sum_{l=1}^{K} \exp(\langle \phi(\mathbf{F}_{ij}), \phi(\mathbf{e}_l)\rangle/\tau)} \cdot \mathbf{p}_k\right] \tag{2}$$

Here, $\mathbf{g}_{ij} = [x, y, \sigma_x, \sigma_y, \rho]$: final Gaussian parameters; $\mu_{\text{grid}}(i,j) = (x,y)$: absolute grid coordinates; $\mathbf{F}_{ij}$: feature vector at position (i,j); $\mathbf{e}_k$: k-th feature prototype; $\mathbf{p}_k = (\sigma_x^k, \sigma_y^k, \rho^k)$: k-th geometric prototype; $\phi(\cdot)$: MLP projection; τ: learnable temperature parameter; K: number of prototypes.

To address potential modality-specific information loss caused by the shared bank, we introduce dual alignment constraints:

Intermodality Self-Consistency Loss ensures consistent geometric representations [4] of the same anatomical structure across different modalities:

$$\mathcal{L}_{\text{self}} = \mathbb{E}_{(I_A, I_B)\sim\mathcal{D}}[\|\mathcal{G}(I_A) - \mathcal{G}(I_B)\|_F^2] \tag{3}$$

This loss enforces geometric consistency between modalities through prototype alignment:

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{(I_A, I_B)\sim\mathcal{D}}\left[\|\mathcal{G}(I_A) - \mathcal{G}(I_{BA})\|_F^2 + \|\mathcal{G}(I_B) - \mathcal{G}(I_{AB})\|_F^2\right] \tag{4}$$

Here, I_A, I_B are paired multi-modal images from distribution $\mathcal{D}$; $\mathcal{G}(I_A), \mathcal{G}(I_B)$ are geometric representations using modality-specific banks; $\mathcal{G}(I_{BA}), \mathcal{G}(I_{AB})$ are cross-modal representations; $\|\cdot\|_F^2$ is the squared Frobenius norm.

Structural Tensor Regularization introduces differential geometry priors as supervision signals:

$$\mathcal{L}_{\text{struct}} = \mathbb{E}_{I\sim\mathcal{D}}\left[\|\mathcal{G}(I_A) - \mathcal{DZ}(I_A)\|_F^2 + \|\mathcal{G}(I_B) - \mathcal{DZ}(I_B)\|_F^2\right] \tag{5}$$

Here, $\mathcal{DZ}(\cdot)$: Diz-Zenzo operator computing structural tensors from image gradients [56]. Final geometric parameters are synthesized through prototype

weighting:

$$\Sigma_{ij} = \sum_{k=1}^{K} \alpha_{ij}^{k} \cdot \mathbf{p}_k \tag{6}$$

The overall geometric optimization objective is:

$$\mathcal{L}_{\text{geom}} = \lambda_1 \mathcal{L}_{\text{self}} + \lambda_2 \mathcal{L}_{\text{align}} + \lambda_3 \mathcal{L}_{\text{struct}} \tag{7}$$

Here, $\lambda_1 = 0.15$, $\lambda_2 = 0.15$, and $\lambda_3 = 0.70$ represent the weighting coefficients for the self-consistency loss, cross-modal alignment loss, and structural regularization loss, respectively.

Through this synergistic design of shared bank and alignment constraints, we maintain cross-modal geometric consistency while avoiding parameter redundancy and training complexity of traditional multi-bank approaches.

Parameter Compression To enhance computational efficiency, we propose a parameter compression strategy based on geometric importance. The significance of each Gaussian unit is measured by its spatial coverage:

$$s_{ij} = \sqrt{|\Sigma_{ij}|} = \sigma_x \sigma_y \sqrt{1 - \rho^2} \tag{8}$$

Based on significance scores, we retain the top N most important Gaussian units, reducing the parameter count from $H \times W$ to N while preserving geometric information integrity (typical setting $N = 1024$).

3.2 Texture-Aware Image Synthesis

After obtaining modality-invariant geometric prototypes, the core challenge in the second stage is how to achieve realistic texture feature mapping while strictly maintaining anatomical structural consistency. Traditional methods often introduce structural perturbations during texture transfer, leading to anatomical distortions in generated images. To address this, we propose a label-adaptive token optimization mechanism and a differentiable Gaussian renderer to achieve precise texture-structure collaborative modeling.

Label-Adaptive Token Refinement Given the source modality image I_{src} and the geometric prototypes $G_{\text{src}} \in \mathbb{R}^{B \times N \times 5}$ extracted from the first stage, the target modality semantic label l_{tgt} is converted into a continuous feature representation through a label embedding layer:

$$\mathbf{f}_{\text{label}} = \text{Embedding}(l_{\text{tgt}}) \in \mathbb{R}^{B \times D_l} \tag{9}$$

Here, B: batch size; N: number of Gaussian prototypes; D_l: label embedding dimension; G_{src}: source modality geometric prototypes; Embedding$(\cdot)$: sinusoidal positional encoding. We innovatively propose a Label-Adaptive Token Refinement Module (LATR) to achieve geometry-aware texture transfer:

$$\Delta G, \mathbf{f}_{\text{texture}} = \text{LATR}(I_{\text{src}}, G_{\text{src}}, \mathbf{f}_{\text{label}}) \tag{10}$$

Here, $\Delta G \in \mathbb{R}^{B\times N\times 5}$ represents the geometry fine-tuning based on target modality characteristics, $\mathbf{f}_{\text{texture}} \in \mathbb{R}^{B\times N\times D_f}$ denotes the texture features of the target modality, and D_f indicates the texture feature dimension.

The adjusted target geometric prototypes are:

$$G_{\text{tgt}} = G_{\text{src}} + \alpha \cdot \Delta G \tag{11}$$

where α is a hyperparameter controlling the magnitude of geometric adjustment. The core innovation of the LATR module lies in utilizing semantic guidance from the target modality to adaptively adjust local geometric details while maintaining the main anatomical structure unchanged, thereby better matching the imaging characteristics of the target modality.

Differentiable Gaussian Renderer Based on the optimized geometric structure G_{tgt} and target texture features $\mathbf{f}_{\text{texture}}$, we construct a complete Gaussian image representation:

$$\mathbf{T}_{\text{tgt}} = [G_{\text{tgt}}|\mathbf{f}_{\text{texture}}] \in \mathbb{R}^{B\times N\times(5+D_f)} \tag{12}$$

To achieve the transformation from sparse tokens to dense images, we design a fully differentiable Gaussian renderer. For any position $x \in \mathbb{R}^2$ in the image plane, the rendering result is obtained through the weighted contribution of all Gaussian units:

$$R(x) = \sum_{k=1}^{N} w_k \cdot \mathcal{N}(x; \mu_k, \Sigma_k) \cdot \mathbf{f}^k_{\text{texture}} \tag{13}$$

where the Gaussian kernel function is defined as:

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{14}$$

Here, $R(x)$ denotes the rendering result at position x, w_k represents the weight of the k-th Gaussian unit automatically computed based on significance, μ_k, Σ_k denote the position and shape parameters of the k-th Gaussian unit, and $\mathbf{f}^k_{\text{texture}}$ represents the texture feature of the k-th Gaussian unit.

This rendering process converts N sparse Gaussian tokens into a complete feature map $R \in \mathbb{R}^{B\times C\times H\times W}$, which is finally passed through a lightweight decoder network to generate the target modality image I_{tgt}.

Multi-Objective Optimization Strategy To balance structural fidelity and texture realism, we adopt a multi-objective optimization strategy:

$$\mathcal{L}_{\text{total}} = \beta_{\text{struct}}\mathcal{L}_{\text{struct}} + \beta_{\text{texture}}\mathcal{L}_{\text{texture}} + \beta_{\text{pixel}}\mathcal{L}_{\text{pixel}} \tag{15}$$

Specifically, the structural consistency loss ensures anatomical accuracy through multi-scale SSIM measurement:

$$\mathcal{L}_{\text{struct}} = 1 - \text{SSIM}_{\text{multi-scale}}(I_{\text{pred}}, I_{\text{GT}}) \tag{16}$$

The texture-aware loss [21] maintains edge sharpness through gradient differences:

$$\mathcal{L}_{\text{texture}} = \mathbb{E}_{I\sim\mathcal{D}}[\|\nabla I_{\text{pred}} - \nabla I_{\text{GT}}\|_1] \tag{17}$$

The pixel-level reconstruction loss ensures global color accuracy:

$$\mathcal{L}_{\text{pixel}} = \mathbb{E}_{I\sim\mathcal{D}}[\|I_{\text{pred}} - I_{\text{GT}}\|_F^2] \tag{18}$$

Here, I_{pred} denotes the predicted target modality image, I_{GT} represents the ground truth target modality image, ∇ denotes the gradient operator, and $\beta_{\text{struct}}, \beta_{\text{texture}}, \beta_{\text{pixel}}$ represent the weight coefficients for each loss term. During training, we employ the following weighting strategy: ($\beta_{\text{struct}} = 0.05$, $\beta_{\text{texture}} = 0.05$, $\beta_{\text{pixel}} = 0.9$).Through this two-stage collaborative design, the Proto-Gaussian framework achieves high-quality modality conversion while maintaining anatomical structural consistency, providing a reliable solution for medical image analysis.

4 Experiment

4.1 Experimental Setup

Experiments are conducted on two public multi-contrast MRI datasets: IXI [28] and BraTS [2]. The IXI dataset contains 100 healthy subjects, and the BraTS dataset contains 150 subjects. To ensure data consistency, all multi-modal volumetric scans from each subject were aligned via affine registration. For intensity normalization, the mean intensity of each volume was first scaled to 1, followed by voxel-wise normalization to the range of $[-1, 1]$. All cross-sectional images were resized to a uniform resolution of 256×256 via zero-padding. For the IXI dataset, we split subjects into 80 for training and 20 for testing, resulting in 4,000 training slices and 1,000 test slices for each translation task. For the BraTS dataset, we split subjects into 120 for training and 30 for testing, yielding 2,880 training slices and 720 test slices. We employ the AdamW optimizer with an initial learning rate of 1×10^{-3} and a cosine annealing scheduling strategy. The mini-batch size is set to 2, and the model is trained for 200 epochs. All experiments are performed on an NVIDIA A100 GPU. For evaluation, we adopt Peak Signal-to-Noise Ratio (PSNR) [42], Structural Similarity Index (SSIM) [49], Learned Perceptual Image Patch Similarity (LPIPS) [58], and Fréchet Inception Distance (FID) as the performance metrics.

4.2 Experimental Results

We compare our method against Pix2pix [22], SAGAN [57], TransUNet [8], ResViT [11], SelfRDB [2], MaskGAN [36], and Contourdiff [9] on IXI (T1↔T2, T1↔PD) and BraTS (T1↔T2) tasks (Tables 1–3). As shown in Figures 3–5, baseline methods suffer from texture blurring, geometric deformation, oversmoothing, or anatomically implausible artifacts. In contrast, our method produces sharper textures and more accurate anatomical structures. On IXI, we

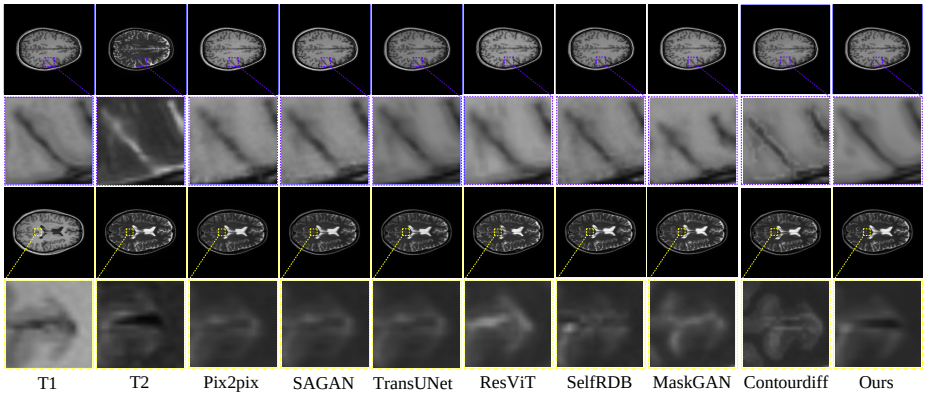

Fig. 3: Qualitative comparison of modality translation performance on the IXI dataset. Top: T2-to-T1 translation results demonstrate the superiority of our method over existing approaches. Bottom: T1-to-T2 translation results further validate the effectiveness of our framework.

Table 1: Quantitative comparison on bidirectional MRI modality translation tasks. Best results are highlighted in bold.

Method	T1 to T2				T2 to T1			
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Pix2pix [22]	27.14	0.9102	0.1259	112.70	26.67	0.9316	0.1029	87.19
SAGAN [57]	28.05	0.9291	0.1031	66.48	27.31	0.9349	0.0901	67.96
TransUNet [8]	27.69	0.8792	0.1415	68.67	27.57	0.9034	0.1300	46.79
ResViT [11]	27.85	0.9283	0.0869	42.93	26.98	0.9236	0.0820	43.76
SelfRDB [2]	29.24	0.9440	0.0766	27.79	28.26	0.9389	0.0798	32.73
MaskGAN [36]	28.81	0.9443	0.0796	20.27	28.00	0.9395	0.0828	37.67
Contourdiff [9]	28.02	0.9257	0.0898	25.85	27.75	0.9243	0.0827	33.28
Ours	**30.57**	**0.9512**	**0.0493**	**17.35**	**29.69**	**0.9546**	**0.059**	**25.38**

outperform the best competitor (SelfRDB) by 1.33–1.43 dB in PSNR, while reducing LPIPS by 26.3–36.4% and improving SSIM by 0.007–0.016. On BraTS, PSNR gains are 1.03–1.48 dB with LPIPS reductions of 32.3–42.0%. These results validate the effectiveness of our Gaussian-based explicit representation in preserving structural consistency and perceptual quality.

4.3 Ablation Studies

We conduct ablation experiments on the T1→T2 task to investigate the loss weight configuration and the Gaussian noise level N. For loss weight configurations (Fig. 6 and Table 4), among all weight configurations, Config. F ($\beta_{\text{struct}} = 0.05$, $\beta_{\text{texture}} = 0.05$, $\beta_{\text{pixel}} = 0.90$) achieves the best performance across all metrics, while Config. A (pure structural loss) yields high SSIM but low PSNR,

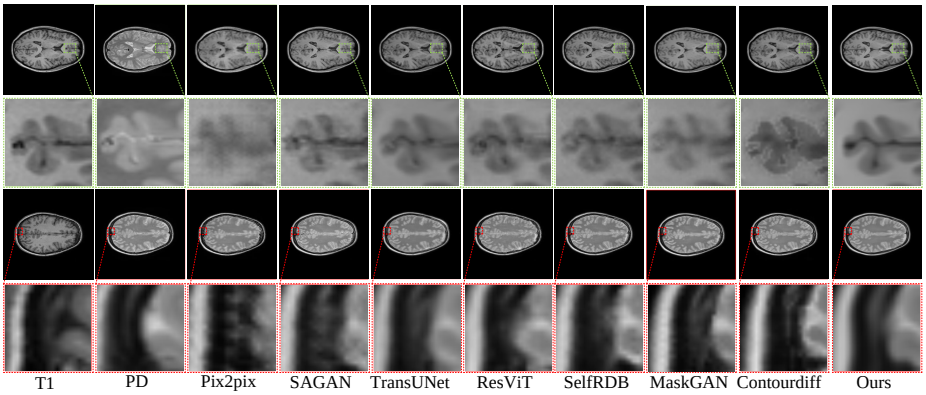

Fig. 4: Qualitative comparison of modality translation performance on the IXI dataset. Top: PD-to-T1 translation results demonstrate the superiority of our method over existing approaches. Bottom: T1-to-PD translation results further validate the effectiveness of our framework.

Table 2: Quantitative comparison on bidirectional MRI modality translation tasks. Best results are highlighted in bold.

Method	T1 to PD				PD to T1			
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Pix2pix [22]	27.01	0.9121	0.1311	143.88	27.56	0.9277	0.1055	90.76
SAGAN [57]	27.57	0.9250	0.0981	65.87	27.97	0.9329	0.0928	56.56
TransUNet [8]	27.91	0.8997	0.0843	69.38	28.10	0.9034	0.0835	34.09
ResViT [11]	27.57	0.9191	0.0872	68.01	27.32	0.9153	0.0871	38.69
SelfRDB [2]	29.03	0.9402	0.0849	58.05	28.66	0.9366	0.0828	40.35
MaskGAN [36]	28.82	0.9370	0.0904	42.37	28.45	0.9322	0.0891	45.02
Contourdiff [9]	27.93	0.9276	0.0886	34.12	27.01	0.9093	0.0985	27.85
Ours	**30.19**	**0.9501**	**0.0501**	**19.88**	**30.02**	**0.9512**	**0.0635**	**26.67**

Config. B (pure texture loss) preserves edges but causes overall degradation, Config. C (pure pixel loss) achieves high PSNR but insufficient structural fidelity, and Config. D (equal weights) leads to mediocre results. Config. E performs close to optimal but slightly underperforms Config. F. For the noise level (Fig. 7 and Table 5), performance peaks at $N = 1024$ and degrades beyond this point. These results demonstrate that a strategy prioritizing pixel-level reconstruction with moderate structural constraints effectively balances pixel accuracy and structural preservation, with $N = 1024$ providing the optimal representation capacity.

4.4 Geometric Consistency

To ensure structural fidelity during texture transformation, we introduce a geometric prior network to learn modality-invariant anatomical representations.

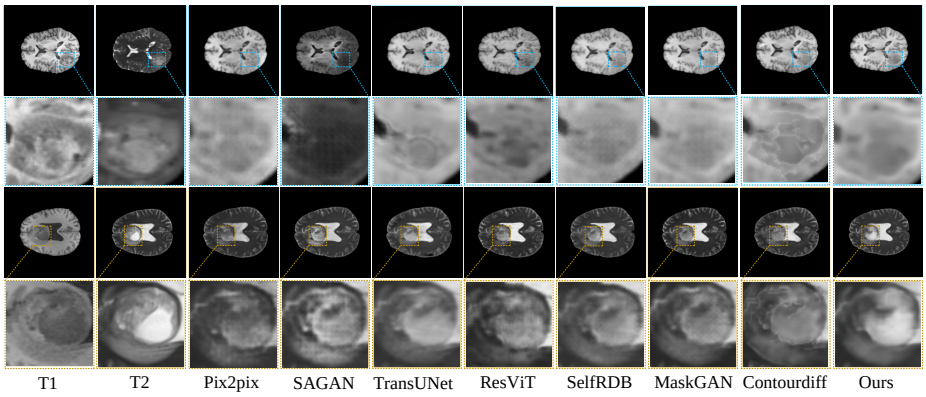

Fig. 5: Qualitative comparison of modality translation performance on the BraTS dataset. Top: T1-to-T2 translation results demonstrate the superiority of our method over existing approaches. Bottom: T2-to-T1 translation results further validate the effectiveness of our framework.

Table 3: Quantitative comparison on BraTs bidirectional modality translation tasks. Best results are highlighted in bold.

Method	T1 to T2				T2 to T1			
	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Pix2pix [22]	27.06	0.9241	0.0913	90.57	26.66	0.9100	0.1126	78.67
SAGAN [57]	27.13	0.9192	0.1014	79.56	26.59	0.9183	0.1063	94.79
TransUNet [8]	27.27	0.9047	0.0823	68.73	26.62	0.8551	0.2052	117.25
ResViT [11]	26.05	0.9017	0.0985	89.59	25.02	0.9075	0.0890	75.21
SelfRDB [2]	27.49	0.9213	0.0975	72.99	26.40	0.9139	0.1133	71.38
MaskGAN [36]	27.38	0.9247	0.0861	29.13	27.14	0.9200	0.0902	**43.14**
Contourdiff [9]	27.23	0.9150	0.0959	22.57	26.60	0.9148	0.1017	48.19
Ours	**28.52**	**0.9428**	**0.0532**	**20.83**	**27.94**	**0.9317**	**0.0784**	43.68

The evaluation of this network follows its design objective by directly verifying its core property—modality invariance of geometric representations. By feeding different modality images (T1, T2, PD) of the same anatomical location into the frozen geometric encoder, we visualize the learned Gaussian parameters (Figure 8).

Each Gaussian prototype is encoded in the HSV color space, where hue represents orientation angle and brightness correlates with scale magnitude. Results demonstrate that despite significant differences in input texture characteristics, the geometric prototypes maintain high consistency in spatial distribution, elliptical shapes, and directional features of key anatomical structures (such as ventricular contours and gray matter boundaries).

In terms of loss function design, we determine weight configurations based on physical significance, prioritizing structural rationality. This configuration

Table 4: Ablation study on loss weight configurations. Best results are highlighted in bold.

Config	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
A: (1.00,0.00,0.00)	28.23	0.9294	0.0636	23.78
B: (0.00,1.00,0.00)	27.19	0.9001	0.0655	27.25
C: (0.00,0.00,1.00)	29.31	0.9290	0.0611	27.31
D: (0.33,0.33,0.33)	29.04	0.9389	0.0576	21.53
E: (0.10,0.10,0.80)	29.72	0.9414	0.0556	18.96
F: (0.05,0.05,0.90)	**30.57**	**0.9512**	**0.0493**	**17.35**

Table 5: Ablation study on Gaussian noise levels (N). Best results are highlighted in bold.

Config	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
N=256	23.63	0.8375	0.1384	159.16
N=512	29.19	0.9268	0.0670	36.76
N=1024	**30.57**	**0.9512**	**0.0493**	**17.35**
N=1536	30.10	0.9480	0.0523	18.75

demonstrates excellent performance in visualization results, simultaneously ensuring anatomical correctness and cross-modal consistency. The successful validation of the geometric prior network provides a reliable structural foundation for subsequent texture transformation stages.

4.5 Computational Efficiency Analysis

We systematically evaluate the computational efficiency and inference cost of different methods on the 256×256 image generation task using an A100 GPU platform. Experimental results demonstrate that our proposed Proto-Gaussian method requires only 6.38 M parameters, which is significantly lower than all compared methods: Pix2pix (54.40 M) [22], SAGAN (7.99 M) [57], TransUNet (27.23 M) [8], ResViT (17.09 M) [11], SelfRDB (8.16 M) [2], MaskGAN (11.75 M), and Contourdiff (16.52 M), making it the most lightweight among all counterparts. Compared with Pix2pix, which has the highest parameter count, our method reduces the parameters by approximately 88.3%. Even compared with the relatively lightweight structures of SAGAN and SelfRDB, our method still reduces the parameters by approximately 20.2% and 21.8%, respectively. In terms of inference cost, our model achieves 134.35 G FLOPs, 15.82 ms latency, and 222.4 MB peak memory. Although these inference costs are higher than some counterparts, the substantial reduction in model parameters offers clear advantages in storage and transmission, which is crucial for deployment in resource-constrained environments. By introducing sparse Gaussian representation and a differentiable rendering mechanism, we achieve high-fidelity image generation

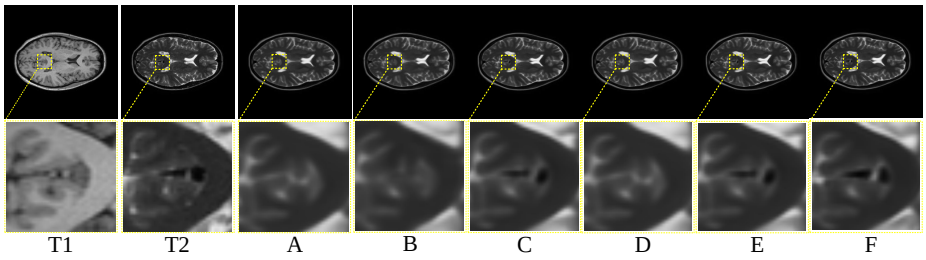

Fig. 6: Ablation study on loss weight configurations for T1→T2 translation. Configuration A yields high SSIM but low PSNR; Configuration B preserves edges but causes overall degradation; Configuration C achieves high PSNR but insufficient structural fidelity; Configuration D leads to mediocre results; Configuration E performs close to optimal but slightly underperforms; while our Configuration F achieves the optimal balance across all evaluation metrics.

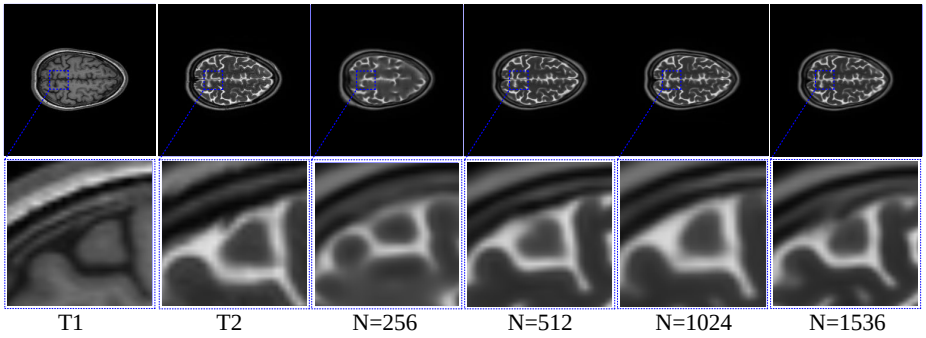

Fig. 7: Ablation study on Gaussian noise levels N for T1→T2 translation.

while significantly reducing model storage and transmission overhead, providing strong support for efficient deployment in resource-constrained scenarios.

4.6 Limitations and Future Work

Despite the competitive results achieved by Proto-Gaussian on 2D axial slices, we acknowledge that MRI data are inherently volumetric, whereas our method processes slices independently without explicitly enforcing inter-slice continuity - a limitation shared by most existing 2D baselines. Our core objective is to establish a Gaussian-based methodology for structure-texture disentanglement, rather than addressing all dimensionality aspects from the outset. Extending this framework to true 3D is a challenging yet promising direction for future work. In addition, many-to-one translation paradigms and systematic evaluation of clinical utility are also important avenues for subsequent investigation. Despite these limitations, we believe our 2D framework offers a novel and solid methodological contribution, as validated on multiple benchmarks and datasets.

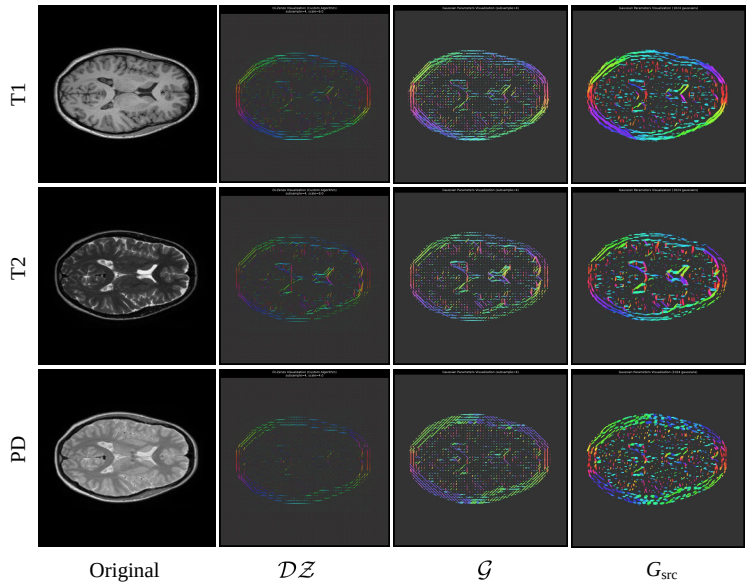

Fig. 8: Visualization of geometric representations across different modalities. For each modality (T1, T2, PD), we display the original image, the structure tensor, the geometric prototype, and the refined geometric structure. The nearly identical geometric visualizations across all modalities validate the modality-invariant property of our geometric prior network.

5 Conclusion

This paper presents Proto-Gaussian, a medical image modality translation framework based on explicit 2D Gaussian representations. Through parametric modeling, our method successfully achieves natural decoupling of anatomical structures and modality-specific textures, effectively addressing the structural distortion problem in cross-modal translation. The proposed two-stage learning pipeline with differentiable Gaussian renderer significantly outperforms state-of-the-art methods across multiple quantitative metrics, without requiring complex task-specific architectures. Extensive experiments demonstrate the superior performance and promising potential of our framework in medical image synthesis applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62472205, 62362051), Academic Leaders Training Program of Jiangxi Province (20232BCJ22001), Key R&D Plan of Jiangxi Province (20232BBE50022).

References

1. Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B.: MedGAN: Medical image translation using GANs. Computerized Medical Imaging and Graphics **79**, 101684 (2020)
2. Arslan, F., Kabas, B., Dalmaz, O., et al.: Self-consistent recursive diffusion bridge for medical image translation. Medical Image Analysis **106**, 103747 (2025)
3. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In: ICCV. pp. 5855–5864 (2021)
4. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
5. Botsch, M., Spernat, M., Kobbelt, L.: Phong splatting. In: Proceedings of the First Eurographics conference on Point-Based Graphics. pp. 25–32 (2004)
6. Chabra, R., Lenssen, J.E., Ilg, E., Straub, T., Lovegrove, S., Newcombe, R.: Deep local shapes: Learning local SDF priors for detailed 3D reconstruction. In: ECCV. pp. 608–625 (2020)
7. Chartsias, A., Joyce, T., Giuffrida, M.V., Tsaftaris, S.A.: Multimodal MR synthesis via modality-invariant latent representation. TMI **37**(3), 803–814 (2017)
8. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
9. Chen, Y., Konz, N., Gu, H., et al.: Contourdiff: Unpaired medical image translation with structural consistency. Machine Learning for Biomedical Imaging **3**, 711–727 (2025)
10. Chung, H., Ye, J.C.: Score-based diffusion models for accelerated MRI. Medical Image Analysis **80**, 102479 (2022)
11. Dalmaz, O., Yurt, M., Çukur, T.: ResViT: Residual vision transformers for multimodal medical image synthesis. TMI **41**(10), 2598–2614 (2022)
12. Dar, S.U.H., Yurt, M., Karacan, L., Erdem, A., Erdem, E., Çukur, T.: Image synthesis in multi-contrast MRI with conditional generative adversarial networks. TMI **38**(10), 2375–2388 (2019)
13. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: NeurIPS. vol. 34, pp. 8780–8794 (2021)
14. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS. vol. 27 (2014)
16. Gu, J., Liu, L., Wang, P., Theobalt, C.: StyleNeRF: A style-based 3D-aware generator for high-resolution image synthesis. arXiv preprint arXiv:2110.08985 (2021)
17. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D medical image segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020)
19. Hu, J., Xia, B., Chen, B., et al.: Gaussiansr: High-fidelity 2d gaussian splatting for arbitrary-scale image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3554–3562 (2025)
20. Hu, S., Lei, B., Wang, S., Wang, Y., Feng, Z., Zhang, Y.: Bidirectional mapping generative adversarial networks for brain MR to PET synthesis. TMI **41**(1), 145–157 (2021)

21. Huber, P.J.: Robust Statistics. John Wiley & Sons (1981)
22. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134 (2017)
23. Jin, C.B., Kim, H., Liu, M., Jung, W., Joo, S., Park, E., Ahn, Y.S., Han, I.H., Lee, J.I., Cui, X.: Deep CT to MR synthesis using paired and unpaired data. Sensors **19**(10), 2361 (2019)
24. Kang, M., Hu, X., Huang, W., Scott, M.R., Wiest, R., Reyes, M.: Dual-stream pyramid registration network. Medical Image Analysis **78**, 102379 (2022)
25. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM TOG **42**(4), 139:1–139:14 (2023)
26. Kim, B., Han, I., Ye, J.C.: DiffuseMorph: Unsupervised deformable image registration using diffusion model. In: ECCV. pp. 347–364 (2022)
27. Lensch, H.P.A., Heidrich, W., Seidel, H.P.: A silhouette-based algorithm for texture registration and stitching. Graphical Models **63**(4), 245–262 (2001)
28. Li, D., Yang, X., Chen, S., et al.: TSMR-Net: A two-stage multimodal medical image registration method via pseudo-image generation and deformable registration. Pattern Recognition Letters **187**, 98–105 (2025)
29. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
30. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3D reconstruction in function space. In: CVPR. pp. 4460–4470 (2019)
31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
32. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM TOG **41**(4), 1–15 (2022)
33. Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., Shen, D.: Medical image synthesis with deep convolutional adversarial networks. IEEE Transactions on Biomedical Engineering **65**(12), 2720–2730 (2018)
34. Niemeyer, M., Geiger, A.: GIRAFFE: Representing scenes as compositional generative neural feature fields. In: CVPR. pp. 11453–11464 (2021)
35. Özbey, M., Dalmaz, O., Dar, S.U.H., Ağdan, S., Çukur, T., Çiçek, Ö.: Unsupervised medical image translation with adversarial diffusion models. TMI **42**(12), 3524–3539 (2023)
36. Phan, V.M.H., Liao, Z., et al., J.W.V.: Structure-preserving synthesis: MaskGAN for unpaired mr-ct translation. In: MICCAI. pp. 56–65. Springer (2023)
37. Ren, L., Pfister, H., Zwicker, M.: Object space EWA surface splatting: A hardware accelerated approach to high quality point rendering. Comput. Graph. Forum **21**(3), 461–470 (2002)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
39. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241 (2015)
40. Rui, S., Chen, L., Tang, Z., Wang, L., Liu, M., Zhang, S., Wang, X.: Multi-modal vision pre-training for medical image analysis. In: CVPR. pp. 5164–5174 (2025)
41. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: GRAF: Generative radiance fields for 3D-aware image synthesis. In: NeurIPS. vol. 33, pp. 20154–20166 (2020)

42. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on Image Processing **15**(2), 430–444 (2006)
43. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265 (2015)
44. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: NeurIPS. vol. 32 (2019)
45. Takikawa, T., Litalien, J., Yin, K., Kreis, K., Loop, C., Nowrouzezahrai, D., Jacobson, A., McGuire, M., Fidler, S.: Neural geometric level of detail: Real-time rendering with implicit 3D shapes. In: CVPR. pp. 11358–11367 (2021)
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS. vol. 30 (2017)
47. Wang, P., Liu, Y., Lin, G., Wang, J., Lu, S., Zhang, W.: Progressively-connected light field network for efficient view synthesis. arXiv preprint arXiv:2207.04465 (2022)
48. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: TransBTS: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–119 (2021)
49. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
50. Wolterink, J.M., Dinkla, A.M., Savenije, M.H.F., Seevinck, P.R., van den Berg, C.A.T., Išgum, I.: Deep MR to CT synthesis using unpaired data. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 14–23 (2017)
51. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, W.T., Tenenbaum, J.B.: MarrNet: 3D shape reconstruction via 2.5D sketches. In: NeurIPS. vol. 30 (2017)
52. Yang, H., Sun, J., Carass, A., Zhao, C., Lee, J., Prince, J.L., Xu, Z.: Unsupervised MR-to-CT synthesis using structure-constrained CycleGAN. TMI **39**(12), 4249–4261 (2020)
53. Yang, Q., Li, N., Zhao, Z., Fan, X., Chang, E.I.C.: MRI cross-modality image-to-image translation. Scientific Reports **10**(1), 3753 (2020)
54. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural radiance fields from one or few images. In: CVPR. pp. 4578–4587 (2021)
55. Yu, B., Zhou, L., Wang, L., Shi, Y., Fripp, J., Bourgeat, P.: EA-GANs: Edge-aware generative adversarial networks for cross-modality MR image synthesis. TMI **38**(7), 1750–1762 (2019)
56. Zenzo, S.D.: A note on the gradient of a multi-image. Computer Vision, Graphics, and Image Processing **33**(1), 116–125 (1986)
57. Zhang, H., Goodfellow, I., Metaxas, D., et al.: Self-attention generative adversarial networks. In: International Conference on Machine Learning. pp. 7354–7363. PMLR (2019)
58. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
59. Zhang, Y., Liu, H., Hu, Q.: TransFuse: Fusing transformers and CNNs for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–24 (2021)
60. Zhang, Z., Yang, L., Zheng, Y.: Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In: CVPR. pp. 9242–9251 (2018)

61. Zhou, H.Y., Guo, J., Zhang, Y., Yu, Y., Wang, L., Yu, Y.: nnFormer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)
62. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017)
63. Zwicker, M., Matusik, W., Durand, F., Pfister, H.: Antialiasing for automultiscopic 3D displays. In: ACM SIGGRAPH 2006 Sketches. pp. 107–es (2006)
64. Zwicker, M., Pfister, H., Baar, J.V., Gross, M.: Surface splatting. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. pp. 371–378 (2001)
65. Zwicker, M., Räsänen, J., Botsch, M., Dachsbacher, C., Pauly, M.: Perspective accurate splatting. In: Proceedings of Graphics Interface. pp. 247–254 (2004)