ResFusion: Medical Image Fusion Driven by Implicit-Forward Diffusion and Time-aware Joint Optimization

Jiahua Xu[1] **Dawei Zhou**[2] **Zhengkun Yu**[1] **Lei Hu**[3] **Lecong Xiong**[1]
Zaiyi Liu[3] **Nannan Wang**[1,*] **Xinbo Gao**[1]

[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China
[2]Faculty of Data Science, City University of Macau, Macao SAR, China
[3]Department of Radiology, Guangdong Provincial People's Hospital, Guangzhou, China

Abstract

Standard training of diffusion models typically relies on a forward diffusion process anchored by ground-truth target images. However, in image fusion tasks, the scarcity of real fused images makes it difficult to formulate a forward process, thereby precluding the standard training. Fortunately, our visual analysis reveals that **image fusion can be effectively modeled without standard training or reliance on the real images.** Nevertheless, the misalignment of input image pairs remains a significant bottleneck for fusion quality. Recognizing that **one-off registration is ill-suited to the progressive generation of diffusion models**, we propose a **unified registration-fusion framework** driven by a time-aware joint optimization mechanism. Specifically, we design a time-aware registration network to progressively optimize registration to guide fusion, while the fused results provide feedback constraints to the registration network at each time step. This mechanism facilitates joint optimization of registration and fusion, significantly improving the quality of fused results. Experimental results show that the proposed method performs excellently on multiple datasets, validating its effectiveness and superiority. Our code is available at `https://github.com/medcx/ResFusion`.

1 Introduction

Multimodal medical image fusion serves a critical role in clinical diagnosis and treatment planning by integrating complementary cross-modal information to synthesize a new modality rich in comprehensive anatomical and lesion details [17]. Recently, Diffusion Models (DMs) have shown great potential in generating high-quality images with their superior modeling capabilities [31, 12]. However, standard training of DMs is typically established upon a forward diffusion process, guided by the distribution of ground-truth target images [27, 28, 23, 26, 29, 42]. In medical image fusion, **the inherent scarcity of real physical fused images renders the construction of such a standard forward process infeasible.** This fundamental mechanism mismatch impedes the seamless deployment of DMs, thereby restricting their efficacy in medical image fusion.

Several approaches have been proposed to solve this problem. One common strategy is to synthesize pseudo-target images, thus indirectly constructing a forward process for training [15, 18, 40]. However, this approach is fundamentally limited, as the generative performance becomes upper-bounded by the capabilities of the pseudo-target images, preventing the model from learning superior fusion representations. Alternatively, unsupervised strategies based on test-time optimization have been proposed, which adapt pre-trained DMs to the fusion task during inference [44, 4]. Nevertheless, since these pre-trained models lack specific priors for the complex correlations between multi-modal

*Corresponding author: nnwang@xidian.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

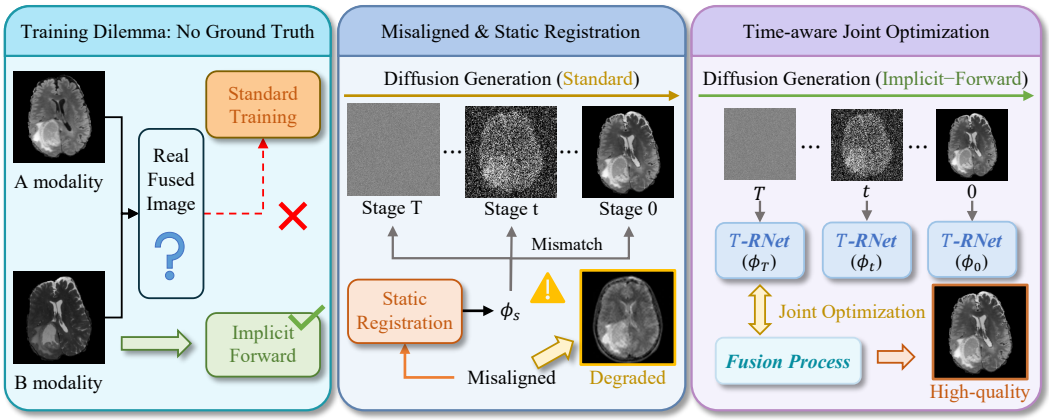

Figure 1: Our research problem and solution schematic. (Left) Standard diffusion training is infeasible due to the lack of real fused images. We introduce an implicit forward diffusion to overcome this limitation. (Middle) Misaligned images lead to structural inconsistencies in the fused results, and a static registration field (ϕ_s) cannot adapt to the progressive generation of DMs. (Right) Our proposed time-aware joint optimization based on implicit forward diffusion: By embedding time information into the network, we achieve time-aware joint optimization of registration and fusion, resulting in high-quality fused results.

medical images, they often struggle to maintain structural consistency of the fused results and **fail to integrate complementary information from different modalities**.

In addition to the limitations of standard training, **spatial misalignment of source images also severely impacts fusion quality**. Most existing fusion algorithms are based on the idealized assumption that the source images are strictly aligned at the pixel level. However, in real clinical scenarios, spatial misalignment is inevitable due to patient movement, varying imaging protocols, and sensor changes [10, 24]. This misalignment can lead to structural inconsistencies and the loss of critical fine-grained lesion details in fused results. To solve this issue, a two-stage 'registration-then-fusion' strategy is adopted [46]. However, these two tasks are inherently related, and such a decoupled strategy fails to exploit the potential for reciprocal guidance, hereby obstructing the optimal integration of multi-modal information. More recently, although some approaches have proposed unified architectures to jointly optimize both tasks [17], they are still constrained by the limitations of static one-off estimation. Thus, they **cannot adapt to the progressive generation of DMs**. This further restricts the application of DMs in medical image fusion and limits the upper bound of fusion tasks.

To address the above challenges, as shown in Fig.1, we propose ResFusion, a unified registration-fusion framework driven by an implicit-forward diffusion model and a time-aware joint optimization mechanism. Firstly, to address the limitation of training caused by the inability to construct the forward process, we demonstrate the feasibility of bypassing the standard training procedure through visual analysis. By constructing an implicit-forward diffusion model, we achieve **effective modeling of image fusion without relying on ground-truth target images**. Secondly, to address the incompatibility between static one-time estimation and DMs, we introduce a time-aware joint optimization mechanism. Specifically, we design a time-aware registration network that leverages cross-modal information guidance and incorporates time information to **overcome the limitations of static estimation**. This allows the registration network to perceive the current time step, enabling it to **strike a balance between global structural correction and local detail refinement**. Based on this, the continuously evolving registration results guide the fusion process, while the fused results provide feedback to constrain the registration network at each time step.

The main contributions are summarized as follows:

- Diffusion models cannot perform standard training without target images, while static estimation fails to adapt to progressive generation. Thus, we propose a novel method ResFusion to address the challenges.
- We propose an implicit forward diffusion model and demonstrate the feasibility of bypassing standard training and performing effective modeling without target images.

- We introduce a time-aware joint optimization mechanism that overcomes the limitations of static estimation by designing a time-aware registration network, thereby achieving joint optimization of both tasks.
- Our method is evaluated on multiple medical image fusion tasks, demonstrating better fused results and superior quantitative metrics scores.

2 Related Work

Deep learning has become the dominant paradigm in multimodal image fusion, evolving from early CNN [9, 8, 43, 38, 19] architectures to DMs [44, 4, 11]. Although DMs excel at generating high-quality images [7], existing methods typically rely on generative priors from pre-trained models [4, 44] or directly utilize "pseudo-fused images" generated by other methods as training guidance [15, 18]. However, these methods are typically based on the ideal of "pixel-level strict alignment," while in real clinical scenarios, source images often exhibit misalignment. Faced with this issue, researchers employ the "registration-then-fusion" two-stage strategy [46, 33]. However, these approaches of separating the tasks transmit the errors introduced during the registration process to the fused result, leading to structural inconsistencies.

To address the limitations of independent processing, unified frameworks have emerged for joint registration and fusion, including ReCoNet [13], UMF-CMGR [35], SuperFusion [32], IMF [36], PAMRFuse+ [45] and BSAFusion [17]. Although unified networks integrate image registration and fusion, they are inherently limited by static one-time estimation. These approaches are incompatible with the progressive generation of diffusion models. Since image content changes during multi-step denoising, static registration cannot provide the necessary adaptive correction, thus limiting the effective application of diffusion models in image fusion tasks.

3 Methodology

3.1 Preliminary

Notice. We denote two different modalities of the same sample and their fused result by the uppercase letters A, B, and F, respectively. Our diffusion model is based on ResShift. Moreover, we let t represent the current timestep of the diffusion process and T represent the total number of timesteps, which is set to 4, according to [41]. We denote a variable following a Gaussian distribution with mean μ and variance σ^2 as $\mathcal{N}(\cdot; \mu, \sigma^2 I)$, where I is the identity matrix.

ResShift. ResShift uses a custom efficient diffusion process to achieve image translation. Specifically, given a pair of images x_0 and y_0, the forward process gradually adds noise to x_0 to obtain a Gaussian distribution with y_0 as the mean. This is represented as follows,

$$q(x_t|x_{t-1}, y_0) = \mathcal{N}\left(x_t; x_{t-1} + \alpha_t e_0, \kappa^2 \alpha_t I\right), \tag{1}$$

where $e_0 = y_0 - x_0$, which represents the residual between the images of source modality and target modality. $\alpha_t = \eta_t - \eta_{t-1}$ for $t > 1$ and $\alpha_1 = \eta_1$, where $\{\eta_t\}_{t=1}^T$ is a sequence that increases monotonically with timestep t and satisfies $\eta_1 \to 0$ and $\eta_T \to 1$. κ is a parameter controlling the variance. Then, we can get the reverse process,

$$p(x_0|y_0) = \int p(x_T|y_0) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y_0) dx_{1:T}, \tag{2}$$

where $p(x_T|y_0) \approx \mathcal{N}(x_T|y_0, \kappa^2 I)$, and $p_\theta(x_{t-1}|x_t, y_0)$ can be expressed as follows:

$$p_\theta(x_{t-1}|x_t, y_0) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, y_0, t), \sum_\theta(x_t, y_0, t)), \tag{3}$$

where $\sum_\theta(x_t, y_0, t) = \kappa^2 \frac{\eta_{t-1}}{\eta_t} \alpha_t I$, and $\mu_\theta(x_t, y_0, t)$ can be expressed as follows:

$$\mu_\theta(x_t, y_0, t) = \frac{\eta_{t-1}}{\eta_t} x_t + \frac{\alpha_t}{\eta_t} \hat{x}_0, \tag{4}$$

where $\hat{x}_0$ is obtained through a network with parameter θ.

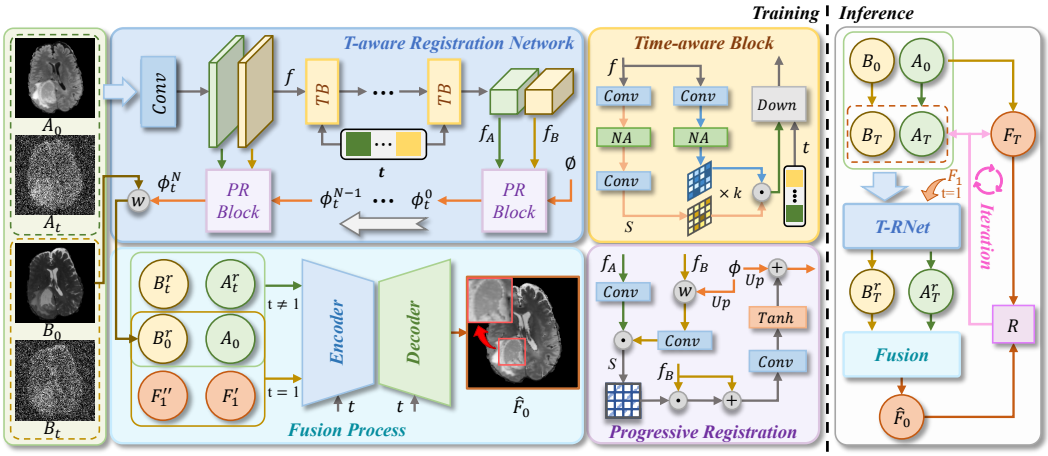

Figure 2: The framework of ResFusion. The framework consists of a time-aware registration network (T-RNet) and a fusion process. Given the source image pair (A_0, B_0) and noisy inputs (A_t, B_t), the T-RNet extracts hierarchical features (f_A, f_B) through time-aware blocks (TB). These features are fed into a progressive registration (PR) block to estimate the deformation field ϕ_t^N in a coarse-to-fine manner. In the fusion process, the aligned image and the fixed image are used as conditional inputs to generate $\hat{F}_0$ through a time-aware guided fusion network. During training, the two networks are jointly optimized. In the inference stage, R represents the reverse process, registration field dynamically evolves to adapt to the generation process, guiding the generation of high-quality fused images.

3.2 Implicit-Forward Image Fusion

Training diffusion models typically relies on a forward process for optimization. However, in fusion tasks, the absence of real fused images prevents the traditional forward process from accessing the necessary ground truth, thereby hindering standard training. To address this issue, we theoretically show that **standard training can be bypassed, enabling effective modeling of fusion tasks.**

In the forward process, we can obtain A_t, B_t, F_t', and F_t'' using Eq. (1):

$$\begin{aligned} A_t &= A_0 + \eta_t(B_0 - A_0) + \kappa\sqrt{\eta_t}\epsilon_t, \\ B_t &= B_0 + \eta_t(A_0 - B_0) + \kappa\sqrt{\eta_t}\epsilon_t, \\ F_t' &= F_0 + \eta_t(A_0 - F_0) + \kappa\sqrt{\eta_t}\epsilon_t, \\ F_t'' &= F_0 + \eta_t(B_0 - F_0) + \kappa\sqrt{\eta_t}\epsilon_t, \end{aligned} \tag{5}$$

where F_0 is the fused image that cannot be obtained. ϵ_t is randomly sampled from $\mathcal{N}(0, I)$.

In the reverse process of fusion task, we can obtain the following representation using Eq. (2):

$$p(F_0|A_0, B_0) = \int p(F_T|A_0, B_0) \prod_{t=1}^{T} p_\theta(F_{t-1}|F_t, A_0, B_0) dF_{1:T}, \tag{6}$$

where $p(F_T|A_0, B_0) \approx \mathcal{N}(F_T|y_0, \kappa^2 I)$, $y_0 \in \{A_0, B_0\}$. And $p_\theta(F_{t-1}|F_t, A_0, B_0)$ can be expressed as:

$$p_\theta(F_{t-1}|F_t, A_0, B_0) = p_\theta(F_{t-1}|F_t, A_t, B_t, A_0, B_0), \tag{7}$$

where A_t, B_t, and F_{t-1} are independent when given F_t, A_0, and B_0. We denote the conditioning information by $obs = (A_t, B_t, A_0, B_0)$.

Then, we denote $p_\theta(F_{t-1}|F_t, obs)$ using Eq. (3):

$$p_\theta(F_{t-1}|F_t, obs) = \mathcal{N}(F_{t-1}; \mu_\theta(F_t, obs, t), \kappa^2 \frac{\eta_{t-1}}{\eta_t}\alpha_t I), \tag{8}$$

where $\mu_\theta(F_t, obs, t)$ can be expressed as follows:

$$\mu_\theta(F_t, obs, t) = \begin{cases} \frac{\eta_{t-1}}{\eta_t} F_t + \frac{\alpha_t}{\eta_t} f_\theta(obs, t), & t > 1, \\ \frac{\eta_{t-1}}{\eta_t} F_t + \frac{\alpha_t}{\eta_t} f_\theta(F_t, A_0, B_0, t), & t = 1, \end{cases} \tag{9}$$

where f_θ is a neural network with parameter θ, and F_1 can be obtained by the reverse process. The theoretical derivation of Eq. (9) can be found in Theorem 1 in Supplementary Material A. We also discuss additional training overhead in Theorem 1.

To ensure that the fused image maintains consistency with the input multi-modal images in terms of spatial structure and anatomical morphology, we introduce a structural consistency loss to constrain the network:

$$\mathcal{L}_{str} = \mathcal{L}_{ssim}(\hat{F}_0, A_0) + \mathcal{L}_{ssim}(\hat{F}_0, B_0), \tag{10}$$

where $\hat{F}_0$ is the prediction of f_θ. This loss measures the structural similarity between the fused image and the source images, effectively suppressing potential structural and geometric distortions introduced during the fusion process.

To further improve the grayscale distribution and contrast of the fused image, we introduce a pixel intensity loss:

$$\mathcal{L}_{pix} = ||\hat{F}_0 - \max(A_0, B_0)||_2^2, \tag{11}$$

where $||\cdot||_2^2$ is the squared Euclidean distance, and $\max(\cdot,\cdot)$ denotes a pixel-wise maximum operator, which selects the maximum intensity value at each spatial location from the two input modality images.

Furthermore, we introduce gradient loss to enhance the ability to preserve edge structure and detail information by constraining the consistency between the fused image and the input multi-modal images in the gradient space:

$$\mathcal{L}_{grad} = ||\nabla\hat{F}_0 - \max(\nabla A_0, \nabla B_0)||_2^2, \tag{12}$$

where ∇ denotes the spatial gradient operator, which computes the image gradients along spatial directions to capture edge and structural information.

Therefore, the fusion loss is as follows:

$$\mathcal{L}_{fuse} = \lambda_1\mathcal{L}_{str} + \lambda_2\mathcal{L}_{pix} + \mathcal{L}_{grad}. \tag{13}$$

To ensure a balanced contribution from both modalities, we set $\lambda_1 = \lambda_2 = 0.5$ in this paper. This setting has been validated to provide a reliable benchmark for integrating various modal features [17, 9].

3.3 T-aware Image Registration and Fusion

T-aware Registration Network. To effectively address the spatial misalignment problem between source images during diffusion, we propose a Time-aware Registration Network r_θ (T-RNet) as shown in Fig.2. Unlike registration methods that rely on static features, our module dynamically adjusts feature extraction based on the current time step t, allowing the registration field to adaptively evolve with the iterative denoising process.

Given the moving image B_0, the fixed image A_0, and their corresponding noisy image (A_t, B_t) at timestep t using Eq. (5), we first employ a parameter-shared convolutional layer for preliminary feature extraction. This step aims to project the input images into a high-dimensional feature space while expanding their channel dimensions. Subsequently, these shallow feature representations (f_A^0, f_B^0) are fed into the Time-aware Block (TB). Within this module, the timestep embedding is injected as conditional information to guide the network in **performing time-aware deep feature extraction**. This process is formulated as follows:

$$f^i = Down(\sum_{j=1}^{k} NA_j(C_j(f^{i-1})) \cdot M^j, t), \tag{14}$$

where $i \in \{1...N\}$ denotes the feature index, N is set to 6, and when $i = 0$, $f^0 = (f_A^0, f_B^0)$. f^{i-1} represents the feature from the preceding layer. $Down$ represents the downsample block, k is the number of parallel branches, which is set to 4. $C_j(\cdot)$ refers to the convolution operation of the j-th branch, while $NA_j(\cdot)$ signifies the normalization and activation layers. M^j denotes the adaptive weight map corresponding to the j-th branch, which determines the pixel-wise contribution of the current branch:

$$M^j = [S(C^1(NA^0(C^0(f^{i-1}))))]_j, \tag{15}$$

where S denotes the softmax function. Subsequently, we extract a set of hierarchical features $\{f_A^i, f_B^i\}_{i=0}^N$ from varying semantic levels. Building upon these multi-scale representations, we devise a Progressive Registration (PR) block to estimate the dense deformation field in a coarse-to-fine manner. The estimation process begins at the coarsest level ($i = N$) to capture large-scale structural displacements. For each subsequent level, we adopt a 'warp-and-refine' strategy. Specifically, the field estimated from the coarser level is first upsampled and utilized to spatially transform (warp) the moving feature f_B^i, achieving preliminary alignment with the fixed feature f_A^i. The registration block then predicts a residual field based on the discrepancy between the warped feature and the target feature. This residual is added to the upsampled field to generate the refined deformation field. This cascade process is formulated as follows:

$$\phi_t^n = Tanh(C(\mathcal{W}(f_B^{N-n}, f_A^{N-n}, \phi_t^{n-1}))) + Up(\phi_t^{n-1}), \tag{16}$$

where $n \in \{1, ...N\}$, which denotes the estimated deformation field at the n-th stage. $Tanh(\cdot)$ is the activation function. $Up(\cdot)$ denotes the upsampling operation using bilinear interpolation. $\mathcal{W}(\cdot, \cdot, \cdot)$ is the spatial warp module, which is formulated as follows:

$$\mathcal{W}(f_B^{N-n}, f_A^{N-n}, \phi_t^{n-1}) = S(C(f_A^{N-n}) \cdot C(\widetilde{f}_B^{N-n})) \cdot \widetilde{f}_B^{N-n} + \widetilde{f}_B^{N-n}, \tag{17}$$

where $\widetilde{f}_B^{N-n} = \mathbf{w}(f_B^{N-n}, \phi_t^{n-1})$, $\mathbf{w}$ represents the Warp operation [14]. For the initial coarsest level ($n = 0$), the warping operation is omitted, and the field is estimated directly from the feature pairs. This cascaded design is crucial for handling large spatial misalignments that coarse-grained layers effectively capture global structural shifts that are often imperceptible at high resolutions due to limited receptive fields. Furthermore, by learning residual field rather than absolute field, the registration task is **significantly reduced in difficulty, resulting in more stable convergence** [6]. Then, we can obtain the aligned image B_0^r:

$$B_0^r = \mathbf{w}(B_0, \phi_t^N). \tag{18}$$

Finally, we optimize the registration network as follows:

$$\mathcal{L}_{reg} = ||B_0^r - B_g||_2^2 + \frac{1}{2}[|\nabla_x \phi_t^N|^2 + |\nabla_y \phi_t^N|^2], \tag{19}$$

where B_g is the aligned ground truth, and the second term regularizes the deformation field ϕ_t^N to ensure spatial continuity and prevent non-physical folding.

Fusion Process. To achieve registration-guided fusion during the generation process, we explicitly introduce the aligned image (A_0, B_0^r) as a conditional input to the denoising network. Specifically, we first obtain the aligned image pairs (A_t^r, B_t^r) at timestep t using Eq. (1):

$$\begin{aligned} A_t^r &= A_0 + \eta_t(B_0^r - A_0) + \kappa\sqrt{\eta_t}\epsilon_t, \\ B_t^r &= B_0^r + \eta_t(A_0 - B_0^r) + \kappa\sqrt{\eta_t}\epsilon_t. \end{aligned} \tag{20}$$

Then, we input it into the prediction network:

$$\hat{F}_0 = \begin{cases} f_\theta(A_t^r, B_t^r, A_0, B_0^r, t), & t > 1, \\ f_\theta(F_t, A_0, B_0^r, t), & t = 1, \end{cases} \tag{21}$$

where f_θ is a SwinUNet [20] based on ResShift [41]. $F_1 \in \{F_1', F_1''\}$, which can be computed by applying the reverse process starting from (F_T', F_T''), respectively. We prove that choosing either one as the starting point for the reverse process yields approximately equal results. The proof can be found in Proof 1 in Supplementary Material A. The fusion loss is redefined as follows:

$$\begin{aligned} \mathcal{L}_{fuse} =& \lambda_1(\mathcal{L}_{ssim}(\hat{F}_0, A_0) + \mathcal{L}_{ssim}(\hat{F}_0, B_g)) + \lambda_2(||\hat{F}_0 - \max(A_0, B_g)||_2^2) \\ &+ ||\nabla\hat{F}_0 - \max(\nabla A_0, \nabla B_g)||_2^2. \end{aligned} \tag{22}$$

Joint Optimization. To achieve joint optimization of r_θ and f_θ, we construct the following optimization objective:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(A_0, B_g)\sim p_{data}, \epsilon\sim\mathcal{N}(0,I)}[\mathcal{L}_{reg} + \mathcal{L}_{fuse}], \tag{23}$$

where p_{data} denotes the joint distribution of aligned multi-modal medical image pairs, and (A_0, B_g) is the training sample sampled from it. B_0 is generated by applying random rigid and non-rigid deformations to B_g. θ^* represents the optimal model parameters after optimization.

This joint optimization fosters **a mutually beneficial mechanism**: the stepwise registration explicitly guides fusion by ensuring precise spatial alignment, while the fusion information conversely acts as a constraint to refine the deformation fields. The details of the overall process are presented in Supplementary Material B.

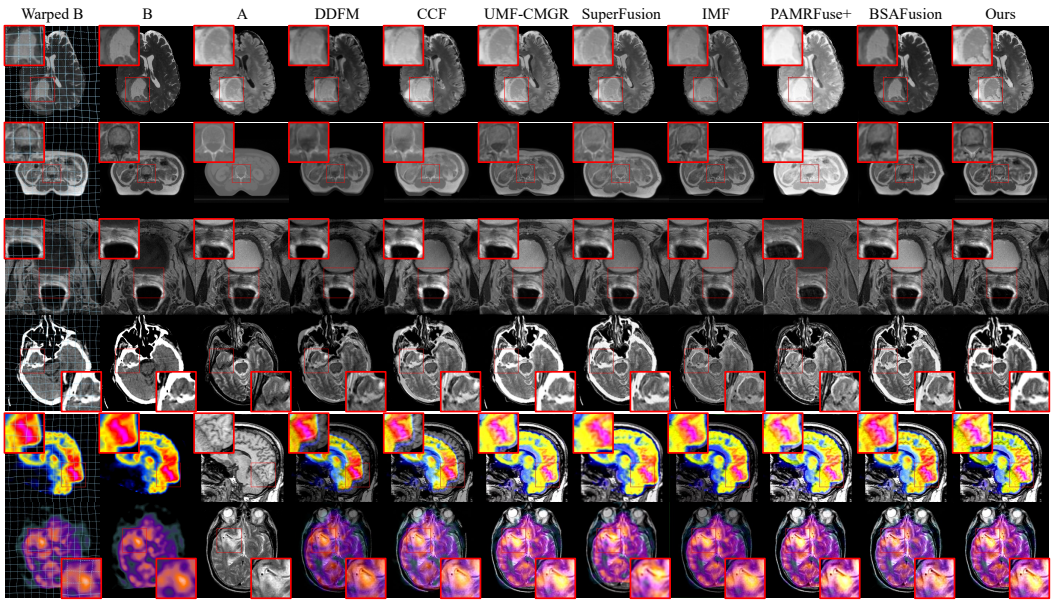

Figure 3: Visual comparison of different approaches. Each row represents a different multimodal fusion task. The first to third columns show the deformed modality B image with spatial misalignment, the corresponding ground truth, and the modality A image, respectively. The fourth to eleventh columns show the fused results generated by different comparison methods and the proposed method, respectively.

4 Implementation Details

4.1 Experimental Datasets

We use four public datasets for comprehensive experiments on six different fusion tasks: (1) BraTs 2023 [2, 3, 22, 1] for T2 & T2-Flair fusion (9,548/2,387 train/eval pairs); (2) SynthRAD 2023 Pelvis [34] for MRI & CT fusion (6,027/1,507 pairs); (3) TCIA Prostate [21] for T1 & T2 fusion (339/83 pairs); and (4) Harvard Whole Brain Atlas [1] for CT/PET/SPECT & MRI fusion (training: 160/245/333; eval: 24). More implementation details are available in Supplementary Material C.

4.2 Comparison Approaches

To comprehensively validate the effectiveness of our proposed method, we select two classes of approaches for comparison. The first class is independent registration and fusion methods, DDFM [44] and CCF [4], for which we use the advanced IMSE [16] approach to perform pre-registration on source images. The second class consists of joint registration and fusion frameworks, including UMF-CMGR [35], SuperFusion [32], IMF [36], PAMRFuse+ [45], and the current state-of-the-art method BSAFusion [17]. All models are trained using the parameter configurations specified in their respective original publications. We quantitatively evaluate the generated results using five standard metrics: Mutual Information (MI) [25], Visual Information Fidelity (VIF) [30], Quality Assessment of Blended Features ($Q_{AB/F}$) [39], Structural Similarity Index Measure (SSIM) [37], and Edge Intensity (EI).

5 Experimental Results

5.1 Comparisons on Image Fusion

Fig.3 shows the visual comparison results, highlighting our method in correcting misalignment while preserving anatomical and lesion details. This is because our time-aware joint optimization mechanism can balance the interaction between registration and fusion within the current time step.

[1] http://www.med.harvard.edu/aanlib/

Table 1: Quantitative comparison of different approaches on BraTs, Pelvis, and Prostate datasets. **Bold** indicates the best and underline indicates the second-best performance.

Dataset	BraTs (T2 & T2-Flair)					Pelvis (MRI & CT)					Prostate (T1 & T2)				
Method	MI	VIF	$Q_{AB/F}$	SSIM	EI	MI	VIF	$Q_{AB/F}$	SSIM	EI	MI	VIF	$Q_{AB/F}$	SSIM	EI
DDFM	2.981	0.380	0.250	1.469	0.107	2.457	0.257	0.147	1.287	0.070	2.165	0.386	0.281	0.865	0.173
CCF	3.073	0.416	0.289	1.520	0.133	2.764	0.353	0.214	1.364	0.088	2.508	0.381	0.313	1.246	0.228
UMF-CMGR	3.731	0.548	0.353	1.531	0.152	3.327	0.425	0.275	1.370	0.104	3.062	0.501	0.358	1.296	0.234
SuperFusion	3.618	0.658	0.420	1.675	0.147	3.363	0.430	0.249	1.365	0.102	3.234	0.671	0.431	1.445	0.228
IMF	3.308	0.504	0.299	1.602	0.111	3.439	0.397	0.267	1.362	0.112	3.830	0.820	0.528	1.508	0.225
PAMRFuse+	3.201	0.408	0.281	1.314	**0.190**	3.012	0.349	0.218	1.180	**0.170**	1.767	0.209	0.177	0.741	0.254
BSAFusion	3.157	0.372	0.211	1.493	0.138	3.119	0.372	0.240	1.348	0.112	3.401	0.672	0.463	1.453	0.263
ResFusion (ours)	**3.881**	**0.842**	**0.578**	**1.718**	0.175	**3.776**	**0.691**	**0.527**	**1.429**	0.144	**3.868**	**0.850**	**0.570**	**1.530**	**0.286**

Table 2: Quantitative comparison of different approaches on Harvard dataset. **Bold** indicates the best and underline indicates the second-best performance.

Dataset	CT & MRI					PET & MRI					SPECT & MRI				
Method	MI	VIF	$Q_{AB/F}$	SSIM	EI	MI	VIF	$Q_{AB/F}$	SSIM	EI	MI	VIF	$Q_{AB/F}$	SSIM	EI
DDFM	2.443	0.264	0.261	1.283	0.204	3.237	0.406	0.374	1.130	0.244	2.446	0.380	0.398	1.349	0.147
CCF	2.340	0.242	0.261	1.285	0.254	2.449	0.399	0.541	1.276	0.337	2.489	0.397	0.495	1.412	0.200
UMF-CMGR	2.994	0.372	0.327	1.396	0.258	3.468	0.655	0.663	1.388	0.400	3.394	0.604	0.640	1.481	0.230
SuperFusion	2.572	0.282	0.248	1.304	0.263	2.159	0.293	0.300	1.174	0.298	2.575	0.365	0.265	1.350	0.161
IMF	2.786	0.307	0.274	1.402	0.200	3.285	0.604	0.446	1.314	0.264	3.095	0.511	0.409	1.466	0.154
PAMRFuse+	3.005	0.465	0.378	1.275	0.262	2.936	0.534	0.659	1.358	**0.457**	2.975	0.568	0.622	1.466	0.261
BSAFusion	2.827	0.396	0.384	1.413	**0.286**	3.579	0.739	0.714	0.713	0.446	3.429	0.733	0.707	0.673	0.268
ResFusion (ours)	**3.086**	**0.486**	**0.385**	**1.419**	0.278	**3.759**	**0.820**	**0.762**	**1.406**	**0.457**	**3.716**	**0.796**	**0.724**	**1.485**	**0.269**

Unlike two-stage approaches with inconsistent structures (e.g., DDFM, CCF), our mechanism ensures high structural consistency (see Supplementary Material D.1 for more results). As shown in Tab.1 and Tab.2, our method achieves better performance on most metrics, particularly MI, VIF, $Q_{AB/F}$, and SSIM, confirming its structural fidelity and information richness. It is worth noting that for the EI metric, our method scores slightly lower than PAMRFuse+ and BSAFusion on certain datasets. However, this does not imply insufficient edge sharpness, but rather reveals the computational limitations of the EI metric itself. EI only measures the intensity of image gradients and is easily misled by high-frequency artifacts, producing a "high-score illusion." As shown in Fig.3, PAMRFuse+ exhibits significant misaligned structures, leading to a high EI. Similarly, BSAFusion shows structural abrupt distortion on CT & MRI fusion task, which increases the gradient value. In contrast, our method generates **more realistic and anatomically consistent edges, resulting in excellent fusion.**

5.2 Downstream Segmentation Analysis

To validate its clinical applicability, we used SAM-Med2D [5] for lesion segmentation on the BraTs (Tab.3). Due to the distribution differences between training images and fused images, the performance of the base model on the fused images is limited; however, some approaches show a visual bias towards T2-Flair modality, resulting in better segmentation results. In contrast, our method achieves the most significant improvement after fine-tuning. This shows that our method **effectively combines lesion localization in T2-Flair with anatomical boundaries in T2, truly achieving cross-modal information fusion.**

Table 3: Comparison of IoU and Dice between basic and fine-tuned SAM-Med2D on BraTs dataset (T2 & T2-Flair).

Method	Basic		Fine-tuned		Improved (%)	
	IoU↑	Dice↑	IoU↑	Dice↑	ΔIoU	ΔDice
T2-warped	0.502	0.648	0.504	0.650	0.40	0.31
T2	0.566	0.702	0.568	0.702	0.35	0.00
T2-Flair	**0.630**	**0.753**	0.668	0.780	6.03	3.59
DDFM	0.588	0.720	0.683	0.792	16.16	10.00
CCF	0.594	0.725	0.658	0.772	10.77	6.48
UMF-CMGR	0.595	0.725	0.680	0.789	14.29	8.83
SuperFusion	0.606	0.734	0.677	0.789	11.72	7.49
IMF	0.608	0.736	0.681	0.793	12.01	7.74
PAMRFuse+	0.618	0.743	0.665	0.779	7.61	4.85
BSAFusion	0.550	0.691	0.648	0.766	17.82	10.85
ResFusion (ours)	0.588	0.719	**0.708**	**0.810**	**20.41**	**12.66**

The visual comparison in Fig.4 provides evidence for these quantitative results. Our method achieves the widest overlap region with the ground truth (yellow area). This indicates that our method enhances the clarity of pathological structures, thereby significantly improving the reliability of

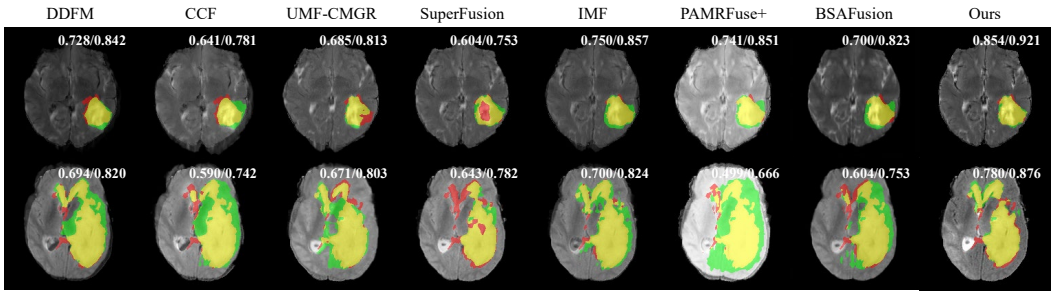

Figure 4: Visual comparison of downstream segmentation results. Red, green, and yellow regions represent the ground truth, prediction, and overlap, respectively. The larger yellow area signifies a higher degree of agreement between the result and the true mask. IoU and Dice values of each image are shown in the corner of images.

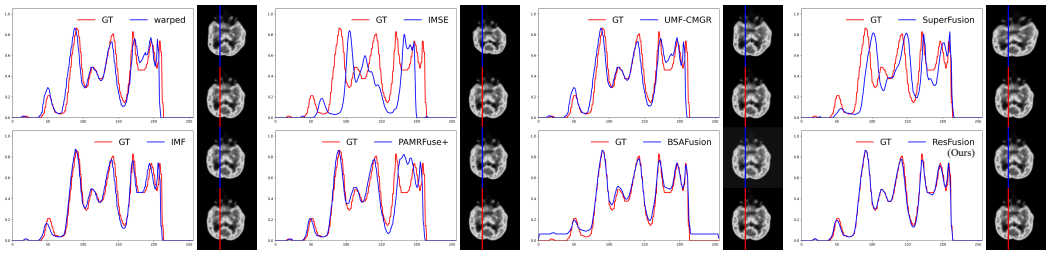

Figure 5: Visualization of the alignment results on PET & MRI task. The blue and red curves represent the pixel overlap at the corresponding blue and red line positions in the aligned image and the ground truth image, respectively. Therefore, a higher degree of overlap between the two curves indicates higher registration accuracy.

subsequent clinical diagnoses. The quantitative metrics in the corner of images also demonstrate the superiority of our method. More segmentation results can be found in Supplementary Material D.2.

5.3 Comparisons on Registration

We visualize alignment performance using pixel intensity distribution maps, referencing [17]. As shown in Fig. 5, it clearly demonstrates that our method achieves excellent registration results. This advantage is particularly prominent in cases where anatomical structures are complex and modality differences are significant (PET & MRI). See Supplementary Material D.3 for more results.

5.4 Ablation Study

We evaluate the contributions of the proposed T-RNet and t-aware mechanism in Table 4. The T-RNet mitigates spatial misalignment between source modalities, while the t-aware mechanism adaptively coordinates registration and fusion across diffusion timesteps. Introducing T-RNet improves MI from 3.360 to 3.847 and $\mathrm{Q_{AB/F}}$ from 0.357 to 0.529, accompanied by consistent gains in VIF and SSIM, demonstrating the importance of spatial alignment for effective cross-modal fusion. Further incorporating the t-aware mechanism yields improvements across all five metrics, achieving the best overall performance. These results highlight the complementary benefits of spatial registration and timestep-adaptive joint optimization in enhancing fusion quality. Further experimental analyses are provided in Supplementary Material E.

Table 4: Ablation study on the effects of the registration network and the t-aware mechanism.

T-RNet	t-aware	MI	VIF	$\mathrm{Q_{AB/F}}$	SSIM	EI
×	×	3.360	0.487	0.357	1.527	0.174
✓	×	3.847	0.788	0.529	1.694	0.174
✓	✓	**3.881**	**0.842**	**0.578**	**1.718**	**0.175**

6 Conclusion

In this paper, we propose a unified registration-fusion architecture based on an implicit forward diffusion model and a time-aware joint optimization mechanism. This novel method effectively addresses the training bottleneck of diffusion models in fusion tasks while eliminating fusion distortions caused by spatial misalignment. Although it has the limitation of slightly increasing training time, the additional overhead is negligible compared to the significant improvement in fusion quality. Extensive evaluations have shown that our method outperforms existing approaches and has been successfully validated in downstream tasks, which is expected to help radiologists make more accurate diagnoses.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant U22A2096, in part by Scientific and Technological Innovation Teams in Shaanxi Province under Grant 2025RS-CXTD-011, in part by the Shaanxi Province Core Technology Research and Development Project under Grant 2024QY2-GJHX-11, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2026SYS-SYSZD-105, in part by the Key Research and Development Program of Shaanxi under Grant 2026JC-ZDGG-42, in part by the Establishment Fund of the State Key Laboratory of Wakefullness/Sleep and Cognition(Chongqing Institute for Brain and Intelligence), in part by the Fundamental Research Funds for the Central Universities under Grant QTZX26138, in part by Science and Technology Development Fund of Macao SAR (0059/2026/ITP1,0054/2025/RIB2), in part by Science and Technology Development Fund of Macao SAR - International Collaborative Research (0001/2025/AIJ), in part by NSFC-FDCT Joint Scientific Research Project Fund (0004/2025/AFJ), in part by National Science Foundation for Young Scientists of China (No.82302130, 62606012).

References

[1] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[2] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, JS Kirby, JB Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection [data set]. the cancer imaging archive, 2017.

[3] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[4] Bing Cao, Xingxin Xu, Pengfei Zhu, Qilong Wang, and Qinghua Hu. Conditional controllable image fusion. *Advances in Neural Information Processing Systems*, 37:120311–120335, 2024.

[5] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

[6] Xinxing Cheng, Tianyang Zhang, Wenqi Lu, Qingjie Meng, Alejandro F Frangi, and Jinming Duan. Sacb-net: Spatial-awareness convolutions for medical image registration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5227–5237, 2025.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[8] Jing Di, Wenqing Guo, Jizhao Liu, Li Ren, and Jing Lian. Ammnet: A multimodal medical image fusion method based on an attention mechanism and mobilenetv3. *Biomedical Signal Processing and Control*, 96:106561, 2024.

[9] Xianming Gu, Lihui Wang, Zeyu Deng, Ying Cao, Xingyu Huang, and Yue-min Zhu. Adaptive spatial and frequency experts fusion network for medical image fusion. *Biomedical Signal Processing and Control*, 96:106478, 2024.

[10] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31(1):8, 2020.

[11] Dan He, Weisheng Li, Guofen Wang, Yuping Huang, and Shiqiang Liu. Dm-fnet: Unified multimodal medical image fusion via diffusion process-trained encoder-decoder. *arXiv preprint arXiv:2506.15218*, 2025.

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[13] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European conference on computer Vision*, pages 539–555. Springer, 2022.

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[15] Yuchan Jie, Yushen Xu, Xiaosong Li, Fuqiang Zhou, Jianming Lv, and Huafeng Li. Fs-diff: Semantic guidance and clarity-aware simultaneous multimodal image fusion and super-resolution. *Information Fusion*, 121:103146, 2025.

[16] Lingke Kong, X Sharon Qi, Qijin Shen, Jiacheng Wang, Jingyi Zhang, Yanle Hu, and Qichao Zhou. Indescribable multi-modal spatial evaluator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9853–9862, 2023.

[17] Huafeng Li, Dayong Su, Qing Cai, and Yafei Zhang. Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4725–4733, 2025.

[18] Mining Li, Ronghao Pei, Tianyou Zheng, Yang Zhang, and Weiwei Fu. Fusiondiff: Multi-focus image fusion using denoising diffusion probabilistic models. *Expert Systems with Applications*, 238:121664, 2024.

[19] Weisheng Li, Yin Zhang, Guofen Wang, Yuping Huang, and Ruyue Li. Dfenet: A dual-branch feature enhanced network integrating transformers and convolutional feature learning for multimodal medical image fusion. *Biomedical Signal Processing and Control*, 80:104402, 2023.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[21] Anant Madabhushi and Michael Feldman. Fused radiology-pathology prostate dataset. *The Cancer Imaging Archive*, 9, 2016.

[22] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[23] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[24] Francisco PM Oliveira and Joao Manuel RS Tavares. Medical image registration: a review. *Computer methods in biomechanics and biomedical engineering*, 17(2):73–93, 2014.

[25] Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics letters*, 38(7):313–315, 2002.

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[28] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[30] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[32] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022.

[33] Linfeng Tang, Qinglong Yan, Xinyu Xiang, Leyuan Fang, and Jiayi Ma. C2rf: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *International Journal of Computer Vision*, 133:5262–5280, 2025.

[34] Adrian Thummerer, Erik van der Bijl, Arthur Galapon Jr, Joost JC Verhoeff, Johannes A Langendijk, Stefan Both, Cornelis (Nico) AT van den Berg, and Matteo Maspero. Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. *Medical physics*, 50(7):4664–4674, 2023.

[35] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *IJCAI*, pages 3508–3515, 2022.

[36] Di Wang, Jinyuan Liu, Long Ma, Risheng Liu, and Xin Fan. Improving misaligned multi-modality image fusion with one-stage progressive dense registration. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):10944–10958, 2024.

[37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[38] Xinyu Xie, Xiaozhi Zhang, Shengcheng Ye, Dongping Xiong, Lijun Ouyang, Bin Yang, Hong Zhou, and Yaping Wan. Mrscfusion: Joint residual swin transformer and multiscale cnn for unsupervised multimodal medical image fusion. *IEEE Transactions on Instrumentation and Measurement*, 72:1–17, 2023.

[39] Costas S Xydeas and Vladimir Petrovic. Objective image fusion performance measure. *Electronics letters*, 36(4):308–309, 2000.

[40] Xunpeng Yi, Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024.

[41] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[43] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5906–5916, 2023.

[44] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8082–8093, 2023.

[45] Yutian Zhong, Jinchuan He, Zhichao Liang, Shuangyang Zhang, Qianjin Feng, Wufan Chen, and Li Qi. Performance of medical image fusion in high-level analysis tasks: A mutual enhancement framework for unaligned pat and mri image fusion. *arXiv e-prints*, pages arXiv–2407, 2024.

[46] Yutian Zhong, Shuangyang Zhang, Zhenyang Liu, Xiaoming Zhang, Zongxin Mo, Yizhe Zhang, Haoyu Hu, Wufan Chen, and Li Qi. Unsupervised fusion of misaligned pat and mri images via mutually reinforcing cross-modality image generation and registration. *IEEE Transactions on Medical Imaging*, 43(5):1702–1714, 2023.

Supplementary Material

A Implicit-Forward Diffusion Model for Image Fusion

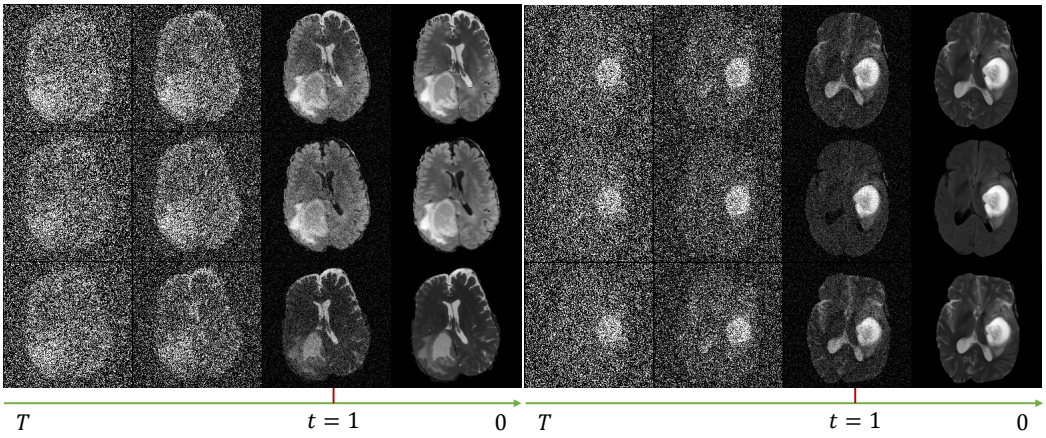

Figure 6: Visualization of different modalities at different time steps. Both images on the left and right are from BraTs. The first to third rows represent the fused image, modality A image, and modality B image, respectively. It can be seen that when $t > 1$, F_t still contains substantial noise, and when $t = 1$, F_1 contains substantial fused information.

Theorem 1. We hypothesize that at $t > 1$, F_t still contains substantial noise, which limits its multi-modal fusion information and consequently renders its influence on F_0 relatively minor, such that it can be effectively substituted by conditional information from A_t and B_t. In contrast, at $t = 1$, F_1 contains substantial fused information, making the contributions of A_1 and B_1 to F_0 relatively minor, since their information can be largely substituted by F_1. Therefore, the conditional probability can be represented as $p_\theta(F_0|F_t, obs)_{t>1} \approx p_\theta(F_0|obs)$ and $p_\theta(F_0|F_t, obs)_{t=1} \approx p_\theta(F_0|F_1, A_0, B_0)$. As shown in Fig. 6, this hypothesis can be intuitively verified through a comparison of the fused image and the multi-modal input at time t. Based on this hypothesis, we can propose the following theory:

$$\begin{aligned} p_\theta(F_{t-1}|F_t, obs) &= \int p_\theta(F_{t-1}, F_0|F_t, obs)dF_0 \\ &= \int p_\theta(F_{t-1}|F_0, F_t, obs) \cdot p_\theta(F_0|F_t, obs)dF_0 \\ &= \int p(F_{t-1}|F_0, F_t) \cdot p_\theta(F_0|F_t, obs)dF_0, \end{aligned} \tag{24}$$

where $p_\theta(F_0|F_t, obs)$ can be expressed as follows:

$$p_\theta(F_0|F_t, obs) \approx \begin{cases} p_\theta(F_0|obs), & t > 1, \\ p_\theta(F_0|F_t, A_0, B_0), & t = 1. \end{cases} \tag{25}$$

Since we employ a neural network f_θ to perform a deterministic estimation of F_0, the posterior distribution is approximated as: $p_\theta(F_0|obs)_{t>1} \approx \delta(f_\theta(obs, t) - F_0)$ and $p_\theta(F_0|F_t, A_0, B_0)_{t=1} \approx \delta(f_\theta(F_t, A_0, B_0, t) - F_0)$. Consequently, the integral over F_0 degenerates into a point mass evaluation at the estimated value:

$$\begin{aligned} \int p(F_{t-1}|F_0, F_t) \cdot p_\theta(F_0|obs)_{t>1}dF_0 &\approx \int p(F_{t-1}|F_0, F_t) \cdot \delta(f_\theta(obs, t) - F_0)dF_0 \\ &= p(F_{t-1}|F_0 = f_\theta(obs, t), F_t), \quad t > 1, \end{aligned} \tag{26}$$

Similarly, we have:

$$\int p(F_{t-1}|F_0, F_t) \cdot p_\theta(F_0|F_t, A_0, B_0)_{t=1}dF_0 \approx p(F_{t-1}|F_0 = f_\theta(F_t, A_0, B_0, t), F_t), \quad t = 1. \tag{27}$$

Therefore, Eq. (4) can be redefined as follows:

$$\mu_\theta = \begin{cases} \frac{\eta_{t-1}}{\eta_t} F_t + \frac{\alpha_t}{\eta_t} f_\theta(obs, t), & t > 1, \\ \frac{\eta_{t-1}}{\eta_t} F_t + \frac{\alpha_t}{\eta_t} f_\theta(F_t, A_0, B_0, t), & t = 1, \end{cases} \tag{28}$$

where F_1 can be computed by applying the reverse process starting from F_T. Since we set $T = 4$ and only introduce the reverse process at $t = 1$ during training, obtaining F_1 brings only marginal computational overhead. Specifically, the static baseline requires 10 minutes per epoch on BraTs, while our method requires only 11 minutes, with even smaller overhead on other datasets. Such a minor increase is negligible compared with the significant improvement in fusion quality.

Proof 1. In the reverse process, we can obtain the following expression based on Eq. 3:

$$p_\theta(F_{t-1}|F_t, A_0, B_0) = \mathcal{N}(F_{t-1}; \mu_\theta, \kappa^2 \frac{\eta_{t-1}}{\eta_t} \alpha_t I). \tag{29}$$

And the difference at the starting points is as follows:

$$\begin{aligned} F'_T - F''_T &= \underbrace{(F_0 + \eta_T(A_0 - F_0) + \kappa\sqrt{\eta_T}\epsilon_T)}_{F'_T} - \underbrace{(F_0 + \eta_T(B_0 - F_0) + \kappa\sqrt{\eta_T}\epsilon_T)}_{F''_T} \\ &\approx (A_0 + \kappa \cdot \epsilon_T) - (B_0 + \kappa \cdot \epsilon_T) \\ &= A_0 - B_0. \end{aligned} \tag{30}$$

Given that both terms share identical variances at the same time step, the stepwise difference in the reverse process reduces to the difference of their means, i.e. $F'_t - F''_t = \mu'_{t-1} - \mu''_{t-1}$. Therefore, we can obtain the following relationship:

$$\begin{aligned} F'_{t-1} - F''_{t-1} &= \mu'_{t-1} - \mu''_{t-1} \\ &= \underbrace{(\frac{\eta_{t-1}}{\eta_t} F'_t + \frac{\alpha_t}{\eta_t} \hat{F}_0)}_{\mu'_{t-1}} - \underbrace{(\frac{\eta_{t-1}}{\eta_t} F''_t + \frac{\alpha_t}{\eta_t} \hat{F}_0)}_{\mu''_{t-1}} \\ &= \frac{\eta_{t-1}}{\eta_t}(F'_t - F''_t). \end{aligned} \tag{31}$$

Based on the above relationship, we obtain the following recurrence relation:

$$\begin{aligned} F'_1 - F''_1 &= \frac{\eta_1}{\eta_2}(F'_2 - F''_2) = \frac{\eta_1}{\eta_2} \cdot \frac{\eta_2}{\eta_3}(F'_3 - F''_3) = \frac{\eta_1}{\eta_2} \cdot \frac{\eta_2}{\eta_3} \cdot \frac{\eta_3}{\eta_4}(F'_4 - F''_4) \\ &\cdots \\ &= \frac{\eta_1}{\eta_2} \cdot \frac{\eta_2}{\eta_3} \cdot \cdots \cdot \frac{\eta_{T-1}}{\eta_T}(F'_T - F''_T) \\ &\approx \frac{\eta_1}{\eta_T}(A_0 - B_0) \to 0 \quad (\text{since } \eta_1 \to 0 \text{ and } \eta_T \to 1). \end{aligned} \tag{32}$$

Given $F_0 = f_\theta(F_1, A_0, B_0)$, and considering the approximation derived above, it follows that the final outputs are equivalent, i.e. $F'_0 \approx F''_0$. To further empirically validate this approximation, we conducted quantitative experiments. The results demonstrate that the relative discrepancy is bounded within the order of 10^{-3}, rendering it negligible.

B Overall Process of ResFusion

To facilitate intuitive understanding, we provide the simple pseudocode as shown in Algorithm 1 (training) and Algorithm 2 (inference).

C Experimental Details

Datasets. BraTs images are center-cropped to the size of 192 × 192, and others are center-cropped to the size of 256 × 256. All images are normalized to [-1, 1]. We obtain the misaligned images by referencing [17].

Algorithm 1 Training process of ResFusion

Input: Two source images A_0 and B_g, number of iterations S, and total diffusion step T.
Output: The optimal model with parameters θ^*.

for $iter = 1$ to S **do**
  Get the misaligned image pairs (A_0, B_0) referencing [17];
  Sampling $t \sim U(\{1, ..., T\})$;
  Sampling $\epsilon_t \sim N(0, I)$;
  Calculate the noisy image (A_t, B_t) using Eq. (5);
  Obtain the deformation field ϕ_t^N by r_θ (T-RNet);
  Calculate the aligned image B_0^r using Eq. (18);
  if $t > 1$ **then**
    Get the aligned image pairs (A_t^r, B_t^r);
  else
    Calculate F_1 by the reverse process, referencing Algorithm 2 ($t > 1$);
  end if
  Obtain the fusion prediction $\hat{F}_0$ using Eq. (21);
  Perform joint optimization for registration and fusion using Eq. (23);
end for
return The optimal network r_{θ^*} and f_{θ^*}.

Algorithm 2 Inference process of ResFusion

Input: Two misaligned source images A_0 and B_0, the optimal model r_{θ^*} and f_{θ^*}, and total diffusion step T.
Output: The fused results F_0.

Sampling F_T from $\mathcal{N}(F_T|A_0, \kappa^2 I)$ or $\mathcal{N}(F_T|B_0, \kappa^2 I)$;
for $t = T$ to 1 **do**
  Calculate the noisy image (A_t, B_t) using Eq. (5);
  Obtain the deformation field ϕ_t^N by r_θ^* (T-RNet);
  Calculate the aligned image B_0^r using Eq. (18);
  if $t > 1$ **then**
    Get the aligned image pairs (A_t^r, B_t^r);
    Calculate $\hat{F}_0 \leftarrow f_\theta^*(A_t^r, B_t^r, A_0, B_0^r, t)$;
    Obtain F_{t-1} using Eq. (8), where μ_θ can be calculated using Eq. (9);
  else
    Calculate $\hat{F}_0 \leftarrow f_\theta^*(F_t, A_0, B_0^r, t)$;
    $F_0 \leftarrow \hat{F}_0$;
  end if
end for
return F_0.

Training. Our framework is based on ResShift [41]. We set the number of channel dimensions to 160, the number of residual blocks to 2, the resolution of the self-attention residual blocks to [48, 24, 12, 6] and [64, 32, 16, 8] for BraTs and others, respectively, and the number of channels per attention head to 32. The channel multipliers were set to [1, 1, 2, 2, 4, 4]. We set the batch size to 4 for BraTs and 2 for others, the total time steps to 4, the window size to 6 for BraTs and 8 for others, and the learning rate to 5.0×10^{-5}, which is then reduced to 2.0×10^{-5} using cosine decay. We trained the model with 100k iterations for BraTs and Pelvis, 10k iterations for others. The diffusion model training is performed on one RTX 4090 GPU.

Evaluation. In the calculation of evaluation metrics, different metrics employ corresponding input ranges. Specifically, the calculation of mutual information (MI) is based on integer grayscale values, with the grayscale range set to [0, 255]; while the remaining evaluation metrics normalize the image intensity to [0,1].

D Additional Experimental Results

D.1 Fusion Visualization

To further demonstrate the effectiveness of our method, we provide additional visual comparisons. As shown in Fig. 7, our results exhibit better structural consistency compared to the comparison approaches and effectively avoid overlapping artifacts caused by registration errors.

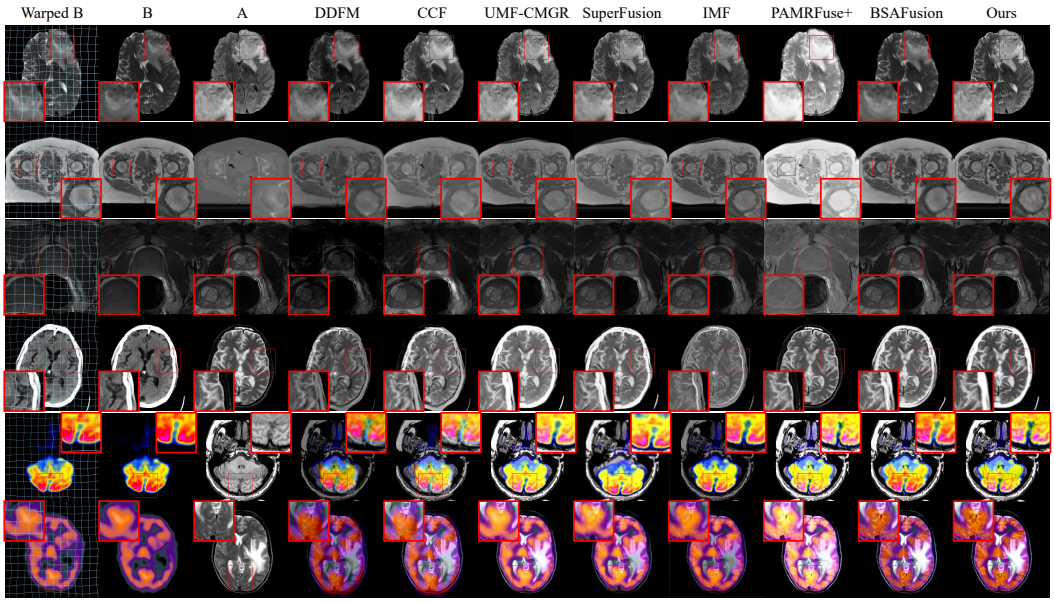

Figure 7: Additional fusion visualization.

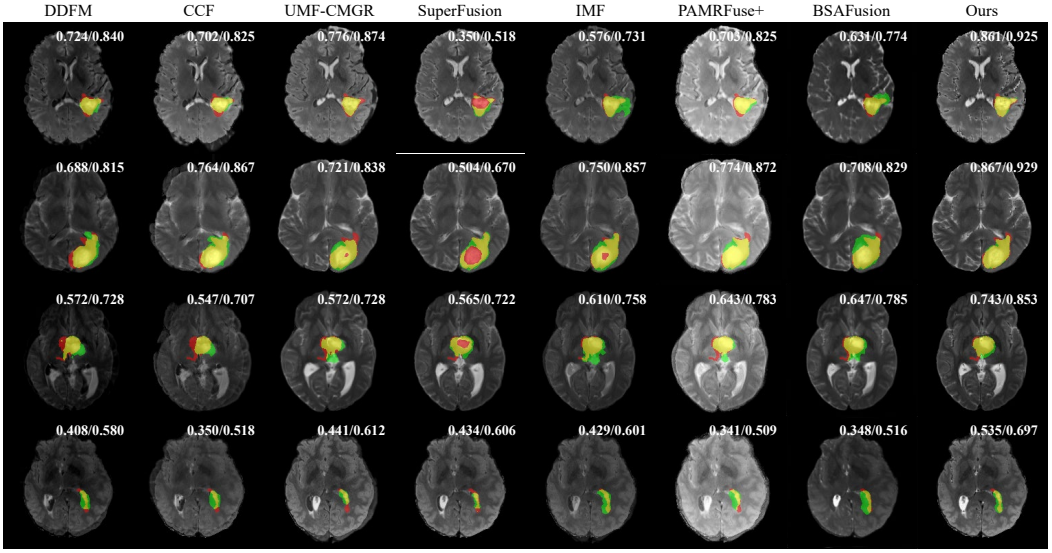

Figure 8: Additional segmentation results on BraTs.

D.2 Segmentation Results

To demonstrate the stability of the segmentation results, we visualize the segmentation results for different lesion sizes, as shown in Fig. 8. The experimental results show that the fused images generated using our method achieve optimal segmentation performance across different lesion sizes.

D.3 Registration Results

We show the registration performance in the remaining five fusion tasks. As shown in Fig. 9, our method achieves excellent performance on different fusion tasks, demonstrating its effectiveness in multi-source data alignment. We also perform quantitative results, as shown in Tab. 5. Our method achieves better performance on most tasks and evaluation metrics. On the PET dataset, although the Normalized Cross Correlation (NCC) score is slightly lower than BSAFusion, our

method still achieves lower Root Mean Square Error (RMSE) values. This result is expected, as NCC primarily measures intensity correlation and is insensitive to small spatial misalignments, while RMSE directly reflects pixel-level alignment accuracy. As shown in Fig. 5, our method exhibits more precise alignment, visually confirming that despite the slightly lower NCC score, our method achieves superior registration quality.

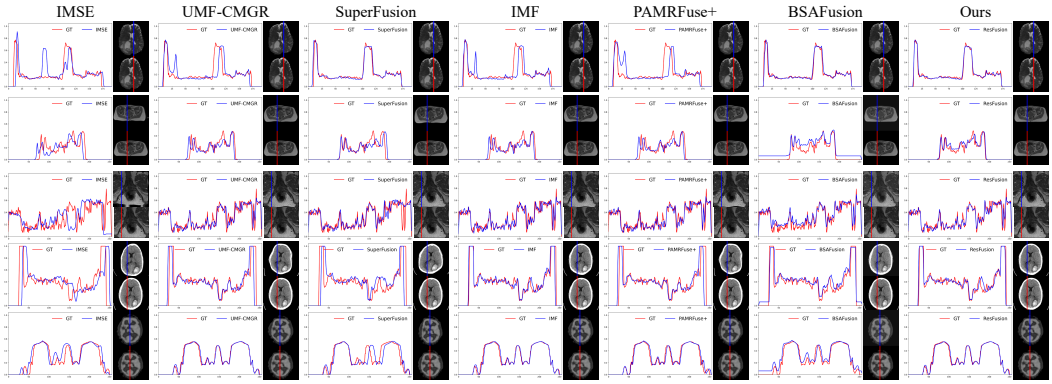

Figure 9: Additional registration visualization on the remaining five fusion tasks. From the first row to the last row, they correspond to BraTs (T2 & T2-Flair), Pelvis (MRI & CT), Prostate (T1 & T2), Harvard (CT & MRI, SPECT & MRI), respectively.

Table 5: Quantitative results of registration. **Bold** indicates the best and underline indicates the second-best performance.

Method	BraTs		Pelvis		Prostate		CT		PET		SPECT	
	NCC↑	RMSE↓	NCC↑	RMSE↓	NCC↑	RMSE↓	NCC↑	RMSE↓	NCC↑	RMSE↓	NCC↑	RMSE↓
IMSE	0.7824	0.1046	0.7147	0.1126	0.5828	0.1161	0.6091	0.2753	0.7195	0.1989	0.8608	0.0855
UMF-CMGR	0.8530	0.0869	0.8793	0.0722	0.8541	0.0695	0.8298	0.1823	0.9449	0.0897	0.9827	0.0305
SuperFusion	0.9599	0.0448	0.7368	0.0999	0.9072	0.0551	0.7334	0.2272	0.7948	0.1797	0.8907	0.0754
IMF	0.8825	0.0773	0.8805	0.0717	0.8701	0.0655	0.9136	0.1306	0.9764	0.0588	0.9824	0.0308
PAMRFuse+	0.8552	0.0862	0.8792	0.0721	0.8405	0.0727	0.8322	0.1805	0.9808	0.0529	0.9821	0.0313
BSAFusion	0.9736	0.0366	0.9191	0.0778	0.8971	0.0705	0.8848	0.1522	**0.9936**	0.0614	0.9745	0.0628
ResFusion (ours)	**0.9817**	**0.0305**	**0.9347**	**0.0528**	**0.9164**	**0.0523**	**0.9138**	**0.1295**	0.9810	**0.0528**	**0.9832**	**0.0301**

E Additional Experimental Analysis

In this section, we provide additional analyses of ResFusion from two perspectives: computational efficiency and registration performance. Specifically, inference time and parameter count characterize the trade-off between computational cost and fusion quality, and comparisons with a registration-only variant further validate the effectiveness of time-aware joint optimization in improving registration accuracy.

E.1 Efficiency Analysis

We compare inference time and parameter count alongside fusion quality to assess the efficiency–performance trade-off of ResFusion.

As shown in Fig. 10, ResFusion achieves the highest VIF on BraTs (0.842) while substantially improving efficiency over existing diffusion-based methods. Although ResFusion is slower and larger than lightweight CNN-based methods, it maintains sub-second inference and delivers substantially higher VIF, demonstrating a favorable trade-off between fusion quality and computational efficiency.

E.2 Effect of Joint Optimization on Registration

To further investigate the contribution of joint optimization to registration performance, we compare ResFusion with a registration-only variant that removes the fusion module while retaining the

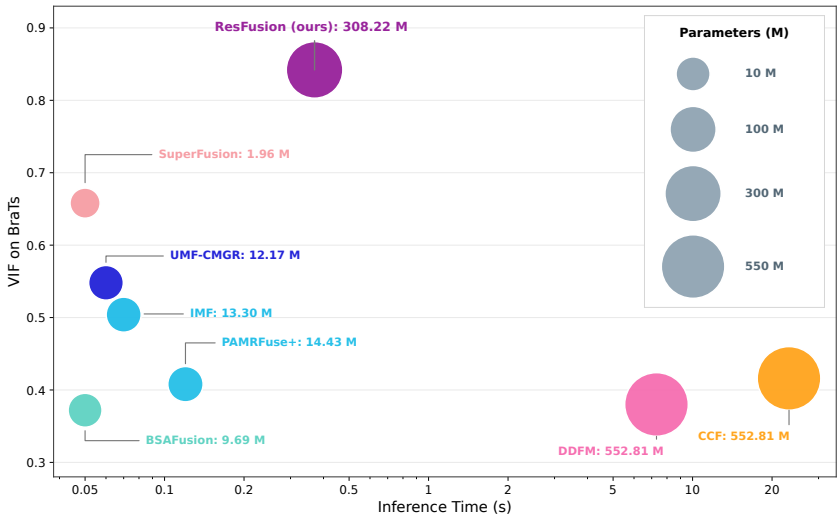

Figure 10: Efficiency–performance comparison of fusion methods. The horizontal axis shows inference time on a logarithmic scale, the vertical axis shows VIF on BraTs, and bubble size represents parameter count. Labels indicate the parameter count of each method.

diffusion-based registration module. As shown in Table 6, ResFusion consistently achieves higher NCC and lower RMSE across all six datasets, demonstrating that fusion-guided joint optimization enhances cross-modal alignment. The improvements are particularly pronounced on BraTs and Pelvis, while consistent gains are also observed on the remaining modality combinations.

Table 6: Comparison of registration performance between the registration-only variant and ResFusion across six datasets. Higher NCC and lower RMSE indicate better registration quality. The best results are highlighted in bold.

Method	BraTs		Pelvis		Prostate		CT & MRI		PET & MRI		SPECT & MRI	
	NCC↑	RMSE↓	NCC↑	RMSE↓	NCC↑	RMSE↓	NCC↑	RMSE↓	NCC↑	RMSE↓	NCC↑	RMSE↓
Registration Only	0.8921	0.0585	0.8636	0.0847	0.9033	0.0542	0.9129	0.1311	0.9758	0.0577	0.9828	0.0302
ResFusion	**0.9817**	**0.0305**	**0.9347**	**0.0528**	**0.9164**	**0.0523**	**0.9138**	**0.1295**	**0.9810**	**0.0528**	**0.9832**	**0.0301**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and Introduction describe these in detail.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Conclusion describe limitations in detail.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The theory and proof are in the supplementary materials A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the model construction in detail and the specific process in the supplementary materials B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We used the public datasets and have open-sourced the code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We explain the training details in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because they are too computationally expensive.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the resources used in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification:

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.