引用格式：罗佳, 刘子枫, 丁熠, 等. 基于扩散的多模态医学图像融合 [J]. 电子科技大学学报, 2026, 55(2): 275-288.
LUO J, LIU Z F, DING Y, et al. Diffusion-based multimodal medical image fusion[J]. Journal of University of Electronic Science and Technology of China, 2026, 55(2): 275-288.

基于扩散的多模态医学图像融合

罗 佳[1]，刘子枫[1]，丁 熠[1, 2*]，卜君健[1]，秦志光[1, 3]

（1. 电子科技大学信息与软件工程学院，成都 610054；2. 宜宾格莱特科技有限公司，宜宾 644000；
3. 西南财经大学计算机与人工智能学院，成都 611130）

摘要：近年来，随着医学影像学技术的持续发展，图像融合技术在医学图像分析中得到了广泛应用。传统的融合技术受限于人工设计的特征提取过程，导致对图像语义信息的理解和匹配精度不高，且无法充分利用多模态图像的信息。通过研究一种基于扩散模型的多模态图像融合方法，利用扩散模型逐步学习多通道图像在潜在空间的联合特征，克服单一端到端网络学习能力有限的问题，生成高质量的融合图像，并对逆向去噪过程针对多模态医学图像融合任务做出了改进。引入了两个模态判别器加强去噪网络对于模态特征的理解，充分利用不同成像模态之间的互补性信息。在 AANLIB 数据集上的实验表明，该方法实现了令人满意的融合结果。

关键词：深度学习；图像融合；扩散模型；多模态
中图分类号：TP391 **文献标志码：**A DOI: 10.12178/1001-0548.2024298

Diffusion-based multimodal medical image fusion

LUO Jia[1], LIU Zifeng[1], DING Yi[1, 2*], BU Junjian[1], and QIN Zhiguang[1, 3]

(1. School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China;
2. Yibin Great Technology Co., Ltd., Yibin 644000, China;
3. School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu 611130, China)

Abstract: In recent years, with the continuous development of medical imaging technology, image fusion techniques have been widely applied in medical image analysis. Traditional fusion methods are limited by manually designed feature extraction processes, resulting in limited accuracy in understanding and matching of image semantic information, and inability to fully utilize the information from multimodal images. A diffusion-based multimodal image fusion method is investigated in this paper. This method progressively learns the joint features of multi-channel images in the latent space using a diffusion model to overcome the limited learning capability of single end-to-end networks. And it generates high-quality fused images and improve the reverse denoising process specifically for the task of multimodal medical image fusion. Two modal discriminators are incorporated to enhance the denoising network's understanding of modality-specific features, fully leveraging the complementary information between different imaging modalities. Experiments on the AANLIB dataset demonstrate that the proposed method achieves satisfactory fusion results.

Key words: deep learning; image fusion; diffusion model; multimodal

随着现代科学技术的发展，影像设备日益变得更加精密和多样化，诸如 MRI、CT、PET 等多模态医学图像在临床应用中变得越来越广泛。同时，医学图像分析技术在现代医学诊疗和研究中扮演着极其重要的角色，特别是在疾病诊断、手术规划、治疗效果评估以及生物医学研究等方面[1]。然而，由于这些成像方式基于不同的物理原理，所获得的图像具有不同的空间分辨率、对比度和解剖结构信息，这种差异性为实现全面精准的医学图像分析带来了挑战。

收稿日期：2024−10−29
基金项目：国家自然科学基金（62076054, 62027827, 62072074, 62372083）；四川省科技计划项目（2022JDJQ0039）
作者简介：罗佳，主要从事深度学习方面的研究。
*通信作者 E-mail: yi.ding@uestc.edu.cn

图像融合是指利用计算机技术将多幅配准后的图像融合成一张图像，能够有效利用多幅图像之间的时空相关性和信息互补性，因此对场景的刻画更加清晰，同时更加具有全面性，无论是人眼识别还是机器检测，都能更清楚地认知图像中所包含的信息[2]。从融合的对象角度来看，图像融合可分为多聚焦图像融合、多曝光图像融合和多模态图像融合等等。其中，从临床医疗的角度看，医学图像是临床诊断的重要依据，临床诊断的准确性需要同时提取多幅不同模态的医学图像信息来保证，因此多模态医学图像融合是多模态图像融合的重要场景[3]。多模态医学图像融合是将来自多种成像设备的图像信息进行有效整合的过程，旨在提取并集中每种模态图像的独特优势，提供更丰富、更全面的解剖和功能信息，进而提升疾病检测的敏感性和特异性，对于早期病变的发现、肿瘤边界识别、神经科学的研究等领域具有极其重大的价值。

针对医学图像的复杂多样特性，融合方法受限于手动特征选取和参数调整，无法满足处理大规模、多模态数据的需求。随着计算机技术演进，深度学习在融合领域展现出显著优势。基于卷积神经网络等深度模型可自动从数据中学习特征及变换关系，减轻对人工特征工程的依赖。对于高维复杂的医学图像，深度学习能有效捕获它们之间的关联性并优化特征表达，从而实现更精确的融合。利用深度学习技术，推动图像融合技术向更高实时性和效能发展，着重全面准确地整合多模态图像之间的互补信息，满足实时手术与诊断需求，确保医疗应用中的快速响应与精准度，是未来研究的一大趋势。

现有方法中，生成模型已被广泛用于模拟融合图像的分布并实现令人满意的融合效果。而近期，扩散模型[3]引起了广泛关注，该模型通过对将噪声损坏的图像恢复到干净图像的扩散过程进行建模来生成高质量的图像。它的生成过程是可解释的，因为它基于去噪扩散来生成图像，从而能够更好地理解图像生成过程。

因此，本文提出了一种去噪扩散图像融合模型，在潜在空间实现多个模态的协同表达，提出了判别器引导的对抗正则化机制，解决多模态特征混淆问题，建立面向医学影像特性的多任务损失函数优化网络的学习方向，实现了多模态融合图像的稳定可控生成。

1 相关技术

1.1 图像融合

图像融合是计算机视觉和图像分领域一个非常重要的研究课题，其拥有广阔的应用领域，包括遥感图像融合[4]、红外可见光图像融合[5]以及临床医学图像融合[6]等。其中，随着医疗保健行业的巨大进步，各种成像传感器的出现改善了临床决策，诊断场景通常需要看到人体的不同深度和结构，通常不能只看到一种模式。因此，来自各种传感器的信息可以组合在一起，以构建新的图像，能够通过其互补信息告诉专家更多信息，因此医学多模态图像融合是一个非常重要的子课题。

在图像融合的各类场景中，主要可分为四类：多时态、多视角、多焦点、多模态融合。融合多种成像模式的主要优点在于每个源图像中的互补信息，以提供更好的场景信息视图，这降低了实现上游任务精确诊断的难度，增强了决策能力，同时降低了存储成本。图像融合可以在空间域或变换域中执行。在空间域中，可以直接处理像素而不是变换系数，根据一些显著性度量直接选择区域或像素，然后与线性或非线性运算合并。在变换域中进行图像融合可以捕获到在空间域中无法捕获的更显著特征。主要而言，融合过程按照抽象层次可以分为三个等级：像素级融合、特征级融合和决策级融合[7]。

早期的研究多为像素级融合方法，该类方法着重于直接针对图像像素点进行分析处理，主要关注图像纹理及结构边缘的处理，并且要求保证融合后图像应保留所有显著信息，且在融合过程中不应引入伪影或不一致，并且应保留移位不变性[8]。在像素级的融合层次上，融合方法可以分为四大类：基于多尺度分解的方法、基于稀疏表示的方法、基于组件替换的方法和基于混合模型的方法[9]，因此，图像融合算法适用于原始或相应的多分辨率图像。为了满足医疗应用的快速发展，像素级融合方法已被广泛用于诊断脑部多模态图像，如 MRI-CT 扫描[10]，PET-MRI[11]和基于 MRI-CT-PET 的肿瘤诊断扫描[12]。像素级融合方法通常比更高级别的技术更容易实现，并且这些方法在一些客观的相似性测量方面使用更微观的信息，在保留图像纹理层面具有显著优势。然而，这类方法对于图像的质量要求很高，要求被融合图像首先被精确地配准，如果图像没有被精确配准，会很大程度上影响融合质量，

并且由于需要逐像素点地处理，通常计算耗时开销大，且融合质量易受噪声影响[13]。

特征级融合算法首先从输入图像中提取语义特征，然后在语义特征层级上进行融合操作。通常，提取语义特征的部分可由一些其他的计算机视觉技术来完成，如语义分割等，用于从输入图像中识别融合主体并使得融合方法聚焦于提升主体在融合后图像中的融合质量。与像素级方法相比，特征级融合方法提取的区域信息量更大，更能准确地关注并区分不同输入图像中有价值的内容，因此，能够更准确地识别不同模态的有用特征，然后融合到一起。相比于像素级的融合方法，特征级融合方法能够有效规避基于像素的算法在对比度、对噪声的敏感性和错误配准方面的缺点，但是其同样存在一定的缺陷，在区域提取过程中，强依赖于分割技术的精确度，分割区域中可能引入一些不良伪影，从而导致空间失真[14]。基于特征级的融合方法，目前也已经有了很多的研究应用，尤其是在医学领域，如脑部 MRI-CT 融合[15]，基于肺部的 FDG PET-T1 MRI 融合[16] 和 Ultrasound-SPECT 融合[17] 等。

决策级融合是最高的抽象层，也称为解释级融合或符号级融合[18]。决策级融合过程会首先从每个源图像中识别和提取特征，然后使用局部分类器进行分类以形成相应的结果，最后应用决策规则来融合它们，旨在加强对对象的共同解释和理解。这一级别的基本挑战在于需要先验知识，由于目标和环境特征的变化，很难获得。因此，算法通常很难实现，在过去的研究中探索较少。最近，随着深度学习的发展，这方面的研究逐渐增多，如文献 [19] 探讨了一种基于中心点遗传算法和模糊逻辑的多模态医学图像融合方法；文献 [20] 提出了在四元小波域中使用稀疏表示分类进行多模态图像融合的方法。

多模态图像融合（multimodal image fusion, MMIF）是数字图像处理与计算机视觉领域中的一个重要研究方向，尤其在医学影像分析、遥感监测、军事侦察和虚拟现实等应用场景中具有广泛而深远的影响。其主要目标是将来自不同成像设备或同一场景、不同时期获取的多种模态图像进行有效整合，以提取并保留每种模态的独特信息，进而生成综合全面且具有高诊断价值或决策支持作用的新图像。

多模态图像融合的过程首先通常是对原始的多模态图像进行必要的预处理，如去噪、增强对比度以及色彩校正等，确保后续融合过程的有效性。然后通过图像配准技术（如前述可变形图像配准），确保不同模态的图像在空间上严格对应，以便进行信息的融合。随后是进行特征提取，从各模态图像中提取能反映各自特性的关键特征，这可能包括但不限于纹理、边缘、强度信息、光谱特性等。特征的选择通常基于各模态图像的优势及其在特定应用领域的价值。最后是根据不同的应用需求和图像特性，采用适宜的图像融合策略，如基于像素级的融合方法（如加权求和、主成分分析、小波变换、稀疏表示等）、基于特征级的融合方法（如融合特征向量、基于深度学习的特征融合）或基于决策级的融合方法（如模糊逻辑、证据理论、D-S 证据理论等）。

在医学影像领域，多模态图像融合用于结合 MRI、CT、PET、超声等多种影像数据，揭示病变的形态学、功能性和生化信息，为疾病诊断、手术规划和疗效评估提供更为详尽准确的依据。传统的多模态医学图像融合方法一般是像素级的融合，着重于对图像中边缘、纹理、梯度等细节信息的处理，缺少对于图像语义特征的理解，无法有效利用不同模态图像间特征的信息互补性。主要的融合方法包括简单的最值化、简单平均、加权平均、独立分量分析（independent component analysis, ICA）、主成分分析（principal component analysis, PCA）、模糊逻辑（fuzzy logic, FL）和云模型（cloud model, CM）等[21]。这些方法仅能在像素级别上对图像中的信息进行分解和处理。如 ICA 利用独立的分量系数对分解后的各部分进行加权融合，而 PCA 则按信息量对各成分分量进行排序，有时会丢弃信息量较少的分量。云模型则是依据源图像生成点云模型，根据隶属度构建相应的融合规则。尽管这些方法在处理图像细节和局部特征上效果显著，但却无法深入理解图像的语义特征，从而限制了对多模态图像丰富信息的充分利用。因此，现有方法虽然在像素级别上能够对图像进行处理和融合，却未能完全挖掘不同模态图像之间的语义信息互补性，这也限制了在医学影像领域对图像全面、高效融合的实现。此外，医学成像技术的快速发展，不同模态间差异大导致融合难度增加，如何平衡各模态信息权重以最大化融合效益，以及如何在保持实时性的同时保证融合质量等也是重大的挑战。

多模态图像融合技术的发展对于推动跨学科交叉研究和解决实际问题具有重要意义，它不仅提升

了单一模态图像所无法提供的丰富信息，而且为现代医学的诊疗及科学研究提供了强大的工具和手段。随着计算能力的不断提升和技术理论的不断创新，多模态图像融合将在未来继续发挥至关重要的作用。

1.2 扩散模型

扩散模型[3]是一种基于概率统计的深度学习框架，可用于生成式建模和图像生成任务。为了实现这一点，模型通常采用变分推断方法，在每个时间步长上估计最优的反向操作，以便将当前的状态一步步拉回到有意义的数据空间中。它通过模拟数据点在连续的时间步长上逐渐从纯噪声状态转变为真实数据分布的过程来实现对复杂数据结构的学习与生成。这一框架的核心思想是构建一个扩散过程，在这个过程中，数据被递增地添加随机噪声，然后训练一个神经网络去学习如何从这种加噪后的状态中逐步还原出原始数据。这个过程受到物理扩散过程的启发，它包含两个主要阶段：正向扩散过程和逆向生成过程，如图 1 所示。

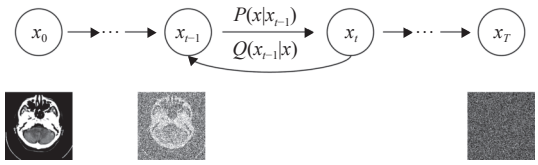


图 1 扩散模型

1）正向扩散过程：正向扩散过程可以视为一个马尔可夫链，该过程以真实的输入数据开始，随着时间步长增加，系统通过不断加入高斯噪声使数据逐渐变为无信息的噪声分布。这一过程通常通过固定迭代次数或设定噪声强度的方式进行，直到达到一个完全不可识别的数据状态。假设$q(x_0)$是真实图像的分布，随机采样一张图像$x\sim q(x_0)$，则正向扩散的过程$q(x_t|x_{t-1})$就是指在当前时间步$t-1$向图像x_{t-1}中添加高斯噪声得到t时刻的图像x_t。由于高斯噪声可被均值μ和方差σ^2唯一定义，那么在每一步添加高斯噪声的过程可以表述如式 (1) 所示：

$$q(x_t|x_{t-1})=\mathcal{N}(x_t|\sqrt{1-\beta_t}x_{t-1},\beta_t x_{t-1}) \tag{1}$$

式中，均值$\mu_t=\sqrt{1-\beta_t}x_{t-1}$，方差$\sigma_t^2=\beta_t$。$\beta$是预定义的差异时间表，用于控制每一时间步混合噪声和图像的比重。

具体到每一步的计算中，首先采样一个高斯噪声$\epsilon\sim\mathcal{N}(0,\mathrm{I})$，随后通过$\beta_t$计算得到$x_t=\sqrt{1-\beta_t}x_{t-1}+\sqrt{\beta_t}\epsilon$。对于策略表$\beta$，常见的 DDPM 策略表有线性时间表、平方时间表、cosine 时间表等。正向扩散的全过程可以表述为时刻$t=1$到时刻$t=T$的马尔科夫链：

$$q(x_{0:T})=q(x_0)\prod_{t=1}^{T}q(x_t|x_{t-1}) \tag{2}$$

正向扩散过程的一个重要特性是x_t可以被看作是原始数据x_0和噪声ϵ的线性组合，可以基于x_0直接采样得到任意时间步t的x_t，该特性可以通过重参数化技巧表示，如式 (3) 所示：

$$x_t\sim q(x_t|x_0)=\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\epsilon=\mathcal{N}(x_t|\sqrt{\bar{\alpha}_t}x_0,1-\bar{\alpha}_t\mathrm{I}) \tag{3}$$

式中，$\bar{\alpha}_t=\prod_1^t\alpha_t$，$\alpha_t=1-\beta_t$。

2）逆向生成过程：经过正向的扩散加噪之后，扩散模型学习逆向过程，即从随机噪声出发，在T时刻采样一个随机高斯噪声，随后逐步去除噪声并重构出接近真实数据分布的样本x_0'。逆向过程的核心操作是如何进行这一去噪过程，这也是整个扩散模型的核心操作。由于在逆向的过程中，$p(x_{t-1}|x_t)$的具体形式是未知的，因此可以通过训练一个生成器网络从任意初始噪声出发，通过一系列时间步长上的优化操作逐步减小噪声并重构出高质量、接近真实数据分布的新样本。因为t时刻的分布x_t是已知的，这个生成器网络的目的就是根据x_t去学习x_{t-1}的概率分布，因此逆向过程可表示为$p_\theta(x_{t-1}|x_t)$，其中θ是网络的参数。

由于正向扩散过程中添加的是高斯噪声，因此在逆向去噪的过程中，为了降低难度，可假设去掉的这部分噪声也是高斯噪声，由于高斯分布可由均值和方差决定，因此$p_\theta(x_{t-1}|x_t)$可表示为：

$$p_\theta(x_{t-1}|x_t)=\mathcal{N}(x_{t-1}|\mu_\theta(x_t,t),\sum_\theta(x_t,t)) \tag{4}$$

式中，均值μ_θ和方差$\sum_\theta(x_t,t)$是根据模型计算得到的，综合所有时间步，逆向去噪过程也可表述为一个马尔科夫链：

$$p_\theta(x_{0:T})=p(x_T)\prod_{t=1}^{T}p_\theta(x_{t-1}|x_t) \tag{5}$$

式中，$p(x_T)$即逆向去噪过程的原始输入，即符合分布$\mathcal{N}(x_T|0,\mathrm{I})$随机采样的高斯噪声；$p_\theta(x_{t-1}|x_t)$是均值和方差未知、需要被预测的高斯分布。

3）目标函数：虽然逆向去噪过程的概率分布$q(x_{t-1}|x_t)$无法直接得到，但是由于已知x_0，可以得

到条件概率分布表述为：

$$q(x_{t-1}|x_t,x_0)=\mathcal{N}(x_{t-1};\widetilde{\mu}(x_t,x_0),\widetilde{\beta}_t\mathbf{I}) \tag{6}$$

根据贝叶斯公式可得：

$$(x_{t-1}|x_t,x_0)=\frac{q(x_t,x_0,x_{t-1})}{q(x_t,x_0)}=q(x_t|x_{t-1},x_0)\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

$$\propto\exp\left(-\frac{1}{2}\left(\left(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\overline{\alpha}_{t-1}}\right)x_{t-1}^2-\left(\frac{2\sqrt{\alpha_t}}{\beta_t}x_t+\frac{2\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_{t-1}}x_0\right)\times x_{t-1}+C(x_t,x_0)\right)\right)$$

$$\propto\mathcal{N}\left(x_{t-1};\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})x_t+\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\overline{\alpha}_{t-1}},\frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{I}\right) \tag{7}$$

式中，$\mu_q(x_t,x_0)=\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})x_t+\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\overline{\alpha}_{t-1}}$是均值，$\sum_q t=\frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\mathbf{I}$是方差。可以看出均值和$x_t$、$x_0$是相关的，但方差与二者是无关的。假设预测的分布也服从正态分布，即：

$$p_\theta(x_{t-1}|x_t)=\mathcal{N}(x_{t-1};\mu_\theta(x_t,t),\sum_q t) \tag{8}$$

通过最小化上述两个分布之间的 KL 散度，可以得到优化的目标函数，表示为：

$$\underset{\theta}{\operatorname{argmin}}D_{\mathrm{KL}}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t))=\underset{\theta}{\operatorname{argmin}}\frac{1}{2\sigma_t^2(t)}\left[\left\|\mu_\theta-\mu_q\right\|_2^2\right]=\underset{\theta}{\operatorname{argmin}}\frac{1}{2\sigma_t^2(t)}\frac{(1-\alpha_t)^2}{(1-\overline{\alpha}_t)\overline{\alpha}_t}\left[\left\|\epsilon_0-\widehat{\epsilon}_\theta\left(\sqrt{\overline{\alpha}_t}x_0+\sqrt{1-\overline{\alpha}_t}\epsilon,t\right)\right\|_2^2\right] \tag{9}$$

最终可以得到逆向去噪过程中去噪网络的优化目标损失函数：

$$\mathcal{L}_{\mathrm{loss}}=\mathbb{E}_{t,x_0,\epsilon}\left\|\epsilon_t-\widehat{\epsilon}_\theta\left(\sqrt{\overline{\alpha}_t}x_0+\sqrt{1-\overline{\alpha}_t}\epsilon,t\right)\right\|^2 \tag{10}$$

式中，ϵ_t是正向扩散过程中 t 时刻添加的高斯噪声，$\widehat{\epsilon}_\theta$是去噪网络预测的噪声。

综上，一旦训练完成，模型就可以从任意的初始噪声中逆向计算出减少噪声的程度，逐步生成与原始数据相似的新的高质量样本。

近年来，尤其是在图像生成领域，扩散概率模型（denoising diffusion probabilistic models, DDPM）[3] 和去噪扩散隐式模型（denoising diffusion implicit models, DDIM）[22] 等的成功应用，使得扩散模型能够在图像分辨率、清晰度和多样性等方面取得突破性进展。此外，在语音合成、视频生成、3D 形状重建等多种任务中，扩散模型也展现出了强大的通用性和潜力。

2　基于扩散模型的多模态图像融合方法

2.1　整体架构

本文的多模态融合方案整体架构如图 2 所示，主要包含判别器训练和扩散过程两个部分。给定两个模态的图像数据，首先训练两个判别器D_{modal1}和D_{modal2}。两个判别器本质上是两个二分类网络，分别用于判断输入图像是否包含属于各自类别的特征，输出是属于各自类别的置信度。训练两个判别器的损失函数采用二分类交叉熵损失，如式 (11) 所示：

$$\mathcal{L}_{\mathrm{BCE}}=-\frac{1}{N}\sum_{i=1}^{N}[y_i\log(p_i)+(1-y_i)\log(1-p_i)] \tag{11}$$

式中，N是一批的样本数量。判别器训练完成后冻结判别器的参数。

然后将两张图像堆叠后送入扩散概率模型。在正向扩散过程中，分为T个时间步长逐步添加噪声；在逆向去噪过程中，利用去噪网络来还原噪声，并对噪声图像逐步去噪。通过正向和反向过程训练扩散模型对潜空间中的多通道图像进行建模，来提取不同模态图像数据的联合潜在数据表达。同时，由于扩散模型是生成式模型，去噪网络只能保证预测的还原噪声符合正向加噪噪声的分布，而无法保证从噪声图像还原出来的图像与源图像的一致性，因此利用两个判别器来引导去噪网络生成合适的噪声，使得还原过程能够生成保证真实性的融合图像。

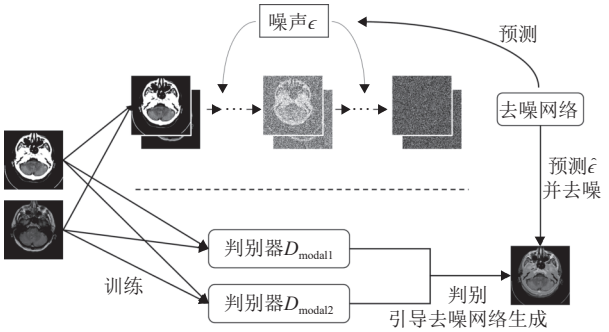


图 2　基于扩散模型的多模态图像融合方法整体架构

2.2 扩散过程

给定一对多模态图像，本文使用了 MRI 和 CT、PET、SPECT 模态图像的组合，其中，MRI 图像以单通道灰度图形式读入，CT、PET、SPECT 图像以三通道 YCrCb 格式读入。为了学习多通道数据在潜空间的联合潜在结构，将两幅图像堆叠在一起后形成四通道输入图像输入扩散网络中，扩散过程如图 3 所示。

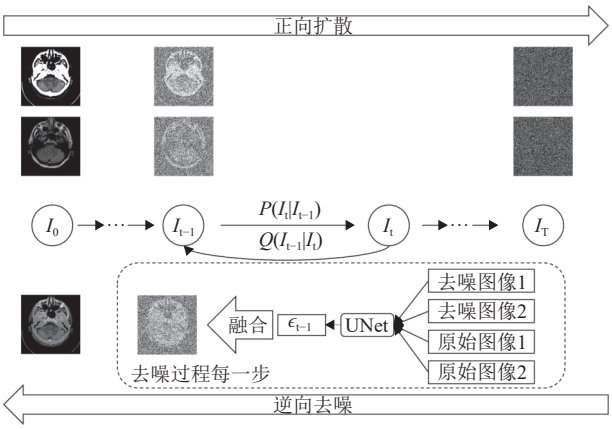


图 3 扩散过程示意图

在正向扩散过程中，以T个时间步逐步添加噪声，在逆向去噪过程中，以T个时间步预测去噪噪声逐步消除噪声并生成融合图像。

具体来说，在正向扩散过程中，每个时间步t生成一个随机高斯噪声来混淆时间步$t-1$时的多通道图像x_{t-1}得到x_t。其中，同扩散概率模型[3]的做法一致，正向扩散过程的一个重要特性是x_t可以被看作是原始数据x_0和噪声ϵ的线性组合，可以通过重参数化技巧，基于x_0直接采样得到任意时间步t的x_t，其过程可以表述为式 (12)，在实际的实践中，α表示采用均匀分布策略得到的线性时间计划表，用于控制噪声和源图像的混淆比例，$\bar{\alpha}_t$表示其时间步t的累积和。

$$x_t \sim q(x_t|x_0) = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{12}$$

因此，可以在逆向去噪的过程中通过上式来获得任意时间步t的噪声图像x_t。具体来说，首先预定义一个线性时间计划表α，然后在逆向去噪过程中的任意时间步t采样一个随机高斯噪声ϵ，通过式 (3) 得到该步的四通道噪声混淆图像，并与原始图像堆叠成八通道图像后输入到去噪网络，去噪网络是一个 UNet[16]，其详细过程如图 4 所示。

去噪网络预测去噪噪声$\widehat{\epsilon}_{t-1}$，利用该去噪噪声对四通道去噪图像进行去噪后将 MRI 图像所在的通道与 YCrCb 三通道中的明度通道进行相加后，得到融合后的去噪图像。其中，针对去噪噪声$\widehat{\epsilon}_{t-1}$计算扩散损失$\mathcal{L}_{\text{diff}}$，以保证去噪网络预测的还原噪声与正向扩散过程的噪声分布一致；针对融合后的去噪图像，计算内容损失$\mathcal{L}_{\text{con}}$和结构相似度损失$\mathcal{L}_{\text{ssim}}$，并使用两个模态的判别器判断融合图像是否包含各自模态特征的置信度，并计算对抗损失$\mathcal{L}_{\text{adv}}$，以此来指导去噪网络生成的噪声满足期望最大化条件，使得融合图像能够保留源图像的内容信息、结构信息以及各自模态的特征信息。详细的损失函数细节在下文进行阐述。

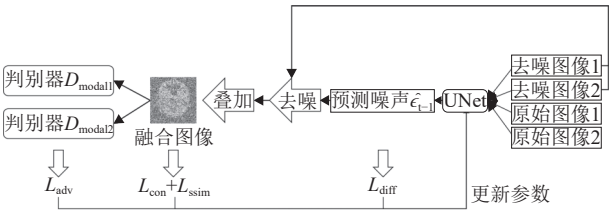


图 4 每个时间步示意图

2.3 损失函数

本方法中去噪网络的损失函数如式 (13) 所示，包含了扩散损失、内容损失、结构相似度损失和对抗损失。

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{ssim}} + \mathcal{L}_{\text{adv}} \tag{13}$$

扩散概率模型的逆扩散过程是为了从纯噪声中恢复原始数据，扩散损失$\mathcal{L}_{\text{diff}}$即扩散概率模型逆向去噪过程中去噪网络的训练损失。该损失项如式 (14) 所示，这个损失函数的思想是让去噪神经网络预测与正向扩散过程中相一致的噪声，然后用这个噪声差异来重构原始数据。

$$\mathcal{L}_{\text{diff}} = \text{MSE}(\epsilon_t - \widehat{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)) \tag{14}$$

但是扩散损失仅约束去噪网络预测的噪声分布，而无法指导扩散过程生成融合后的图像，因此需要引入内容损失作为深监督项，来引导逆向去噪过程中去噪网络预测有利于生成融合结果的噪声分布，使得还原噪声满足融合任务的期望最大化条件。内容损失$\mathcal{L}_{\text{con}}$如式 (15) 所示，其由基于像素值的均方差损失和基于像素值梯度的梯度损失组合而成。基于像素值的均方差损失保证了融合后图像内容的真实性，基于像素值梯度的梯度损失保证了融合后图像的纹理结构与真实图像的一致性。

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{grad}} \tag{15}$$

式中，$\mathcal{L}_{\text{mse}}=\frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}((I_{\text{fused}}-I_{\text{modal1}})^2+(I_{\text{fused}}-I_{\text{modal2}})^2)$，$\mathcal{L}_{\text{grad}}=\frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}((\nabla I_{\text{fused}}-\nabla I_{\text{modal1}})^2+(\nabla I_{\text{fused}}-\nabla I_{\text{modal2}})^2)$。

结构相似度损失$\mathcal{L}_{\text{ssim}}$用于模拟人类的视觉感知损失。由于在实际的医学诊断领域，大多数医生根据成像的局部亮度、对比度、结构差异等进行主观的判断；同时，由于 MRI 图像包含更清晰的结构信息，故引入融合图像与 MRI 图像的结构损失来矫正融合图像的结构信息，使得融合图像的结构更符合人类医生的视觉感知损失。结构相似度损失$\mathcal{L}_{\text{ssim}}$的计算表达式如式 (16) 所示，其中 SSIM 的计算方法在下文介绍。

$$\mathcal{L}_{\text{ssim}}=1-\text{SSIM}(I_{\text{fused}},I_{\text{MRI}}) \tag{16}$$

为了保证融合后图像能够有效保留多个模态的特征，本文引入了对抗损失$\mathcal{L}_{\text{adv}}$，如式 (17) 所示。该损失包含两个判别器D_{modal1}和D_{modal2}，用于判别融合后的图像I_{fused}是否包含各自模态的特征。这两个判别器已在两个模态的数据上完成预训练，用来区分输入数据是否属于本模态，并输出一个属于本模态的概率。

$$\mathcal{L}_{\text{adv}}=(1-D_{\text{modal1}}(I_{\text{fused}}))^2+(1-D_{\text{modal2}}(I_{\text{fused}}))^2 \tag{17}$$

3　实验

3.1　实验环境及配置

本文的实验均基于 Python 的 PyTorch 深度学习库实现，实验环境配置为一台配备英特尔 Gold 6338 Xeon CPU、256 GB 内存和 Nvidia Tesla a100 GPU 的 PC。

扩散过程的时间总步数T设为 1 000 步，时间计划表α使用线性时间计划表，并使用学习率为 0.000 1 的 Adam 优化器对去噪网络进行训练。

3.2　图像融合数据集

为了评估本文提出的多模态医学图像融合方法的有效性，本文与多个对比方法在多个模态的图像上进行了对比实验，并根据多个指标评估了本文方法的有效性，对比方法及评估指标将在后文介绍。本节介绍实验所使用的 AANLIB 数据集。

AANLIB 数据集是一个公开的多模态医学影像数据库，由哈佛大学医学院建立，主要用于教育和研究的目的，可免费下载。该数据库中提供了丰富的脑部解剖以及血流图像，用于研究脑部功能结构以及与年龄的相关性变化。还提供了多种脑血管疾病的图像，对中风等脑血管疾病的原因、症状、诊断和治疗提供了参考和研究依据。此外，该数据库中还提供了多种脑部肿瘤的图像，覆盖不同的肿瘤类型、位置和大小，以及不同的治疗效果。对于如阿尔茨海默病、亨廷顿病、运动神经元病等多种退行性疾病的脑部图像也有涉及。另外还包括了一些炎症或感染性疾病的脑部图像，可用于研究多发性硬化、莱姆脑病、克雅氏病等的表现、诊断和治疗。

对于多模态医学图像融合任务而言，该数据库中主要包含的脑图谱数据，提供了同一位置、同一时间、同一视角下 MRI 和 CT、PET、SPECT 3 种不同模态下各自一一对应的脑部影像。因此，本文在 CT、PET、SPECT 3 种不同模态上进行了融合实验，分别与各自对应的 MRI 图像进行融合，并进行了客观以及主观分析，以此来评估本文提出的多模态融合方法的有效性。

3.3　对比方法

为了验证本文提出的多模态医学图像融合方法的优越性，采用了 5 种不同的对比方法进行综合比较，这些方法分别是 SDDGAN[23]、U2Fusion[24]、MSRPAN[25]、Dif-fusion[26] 以及 MMIF-DDFM[7]。

SDDGAN[23] 是一种基于生成对抗网络的端到端融合模型，用于在多模态图像融合中获得更具语义一致性的图像。特别地，该方法中设计了一个信息量判别块来指导融合过程，对于每个源图像，该块确定保留每个语义对象特征的权重。通过这种方式，生成器学习通过不同的权重融合各种语义对象，以保留它们的特性。此外，双重鉴别器被用来识别融合图像中不同模态信息的分布，每个鉴别器对融合图像中不同语义对象的某一模态起作用，以保留和增强它们的模态特性。该方法生成的融合图像更具信息性，红外图像中的热辐射和可见图像的纹理细节都可以得到很好的保留。

U2Fusion[24] 是一个统一的无监督端到端融合框架，适用于解决包括多模态、多曝光、多焦点等多种任务。该方法使用特征提取和信息测量来自动估计相应源图像的重要性，并提出了自适应信息保存度来保持融合结果和源图像之间的自适应相似性。方法通过联合学习多个模态的图像特征，并在多层次上进行融合，以产生更丰富和一致的信息并优化图像特征的表达，以提高融合后图像的质量和信息完整性。

MSRPAN[25] 采用了多尺度注意力机制，该网络由一个特征提取器、一个融合器和一个重构器组

成。特征提取器用于提取不同尺度的特征，融合器对特征进行加权融合，重构器输出最终融合图像。该方法结合了残差网络的思想，以更有效地捕获医学图像中的关键特征。通过建立金字塔式的注意力机制，能够在不同尺度上聚焦于图像中的重要区域，从而提高融合质量。

Dif-fusion[26] 使用了扩散模型来完成图像融合任务。与现有的将多通道输入数据转换为单通道数据的方法不同，该方法在前向扩散和反向去噪过程的潜空间中，利用去噪网络创建了多通道输入数据的分布，并将多通道扩散特征输入到多通道融合模块中，直接生成三通道融合图像，以此来增强多源信息聚合的能力和颜色的保真度。同时，为了保留纹理信息和强度信息，该方法使用了信息的多通道梯度损失和强度损失。

MMIF-DDFM[7] 是一种基于去噪扩散概率模型的融合方法，其核心在于利用强大的先验知识解决基于 GAN 方法中的不稳定训练和缺乏可解释性问题。该方法将融合任务划分成无条件生成子问题和最大似然子问题，以具有潜变量的层次贝叶斯方式建模，并通过期望最大化算法进行推断。通过将推断解决方案集成到扩散采样迭代中，该方法可以生成具有自然图像生成先验和源图像的跨模态信息的高质量融合图像。该方法只需要一个无条件的预训练生成模型，而无需进行任何微调。

为了保证实验的公正性，本文在相同的实验条件下对这些方法进行评估。在比较实验中，本文将重点关注多模态医学图像融合结果的视觉质量、信息保留能力以及在特定医学任务中的性能表现。通过综合比较这 5 种方法，本文将更全面地评估它们在多模态医学图像融合任务中的性能差异。

3.4 评估指标

由于客观的量化指标并不总是能反映人眼对图像质量的感知，通常采用主客观结合的方式来评估图像融合方法的优越性。在定量实验中，本文采用了标准差（SD）、峰值信噪比（PSNR）、互信息（MI）、视觉信息保真度（VIF）、熵（EN）、Qabf 和结构相似性指数度量（SSIM）7 个指标来综合评估融合结果。

标准差（SD）在图像融合任务中通常被用来衡量图像的对比度。在图像中，像素值的标准差越大，表示图像的灰度级别变化越剧烈，图像的对比度越高。而标准差较小则表示图像的灰度级别变化较为平缓，对比度较低。在图像融合中，对比度的提高通常被认为有助于更好地展示图像的细节和特征。SD 的计算方式如式 (18) 所示：

$$SD = \sqrt{\sum_{i=1}^{H}\sum_{j=1}^{W} I_{\text{fused}}(i,j) - \mu} \tag{18}$$

式中，μ为融合图像I_{fused}的像素值的均值。

峰值信噪比（PSNR）是一种基于 MSE 改进的评估指标，主要衡量图像的保真度，即融合后的图像与原始图像之间的相似度，其单位是 dB。PSNR 的数值越高，表示融合后的图像与原始图像之间的差异越小，图像质量越高。通常，PSNR 的值在 60 以上时，认为融合图像的质量好，保真度高；而在低于 30 时被认为与源图像差异较大，融合质量较差。PSNR 的计算依赖融合图像的像素值最大值MAX和两张图像像素值间的均方误差MSE，具体的计算方式如式 (19) 所示：

$$\text{PSNR} = 10\log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right) \tag{19}$$

互信息（mutual information, MI）在图像融合任务中衡量两个图像之间的信息关联性。具体而言，互信息衡量的是在已知一个图像的情况下，另一个图像的不确定性的减小程度。如果两个图像的像素在相同位置上具有较高的相似性，那么它们之间的互信息就会较高，表示它们之间有更强的关联性。MI 的计算依赖像素值的联合概率分布和边缘概率分布，其计算方式如式 (20) 所示：

$$\text{MI}(X,F) = \sum_{x\in X}\sum_{f\in F} P(x,y)\log\left(\frac{P(x,f)}{P(x)P(f)}\right) \tag{20}$$

式中，$P(x,f)$是输入变量 X 和 F 的联合概率分布，$P(x)$、$P(f)$分别为输入变量 X 和 F 的边缘概率分布。

视觉信息保真度（VIF）在图像融合任务中旨在反映图像融合结果与原始图像之间的信息损失程度。具体而言，VIF 通过比较原始图像和融合图像之间的结构信息来评估融合的保真度。VIF 考虑了两个方面的信息：局部结构和全局结构。它量化了融合图像中的结构信息与原始图像中对应结构信息之间的相似度。VIF 的计算涉及局部和全局的空间域变量，以及在不同频率下对结构信息的处理。其数学表达式如式 (21) 所示：

$$\text{VIF} = c_1 \frac{2\mu_x\mu_y + c_2}{\mu_x^2 + \mu_y^2 + c_2}\frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}$$

$$c_1 = (0.01\max_{x,y}(\mu_x,\mu_y))^2$$

$$c_2 = (0.03\max_{x,y}(\sigma_x,\sigma_y))^2$$

$$c_3 = \frac{c_2}{2} \tag{21}$$

式中，μ_x和μ_y分别是原始图像和融合图像的像素值的局部均值；σ_x和σ_y分别是原始图像和融合图像的像素值的局部标准差，σ_{xy}是原始图像和融合图像的局部协方差。VIF 的取值范围在 0～1，越接近 1 表示融合图像与原始图像之间的结构信息相似度越高，保真度越好。

熵（EN）在图像融合中被用于评估融合结果的信息丰富程度。熵越高，图像包含的信息越复杂，反之越低则表示图像信息相对简单和可预测。在图像融合中，如果融合后的图像保留了更多的细节和纹理，使得灰度级别分布更加分散，那么图像的熵就会相对较高。相反，如果融合导致信息的损失，使得灰度级别分布更加集中，那么图像的熵就会相对较低。其计算公式如式 (22) 所示：

$$\mathrm{EN} = -\sum_{i=1}^{N} p_i \log(p_i) \tag{22}$$

式中，p_i表示图像中灰度级别i的概率，N是图像的灰度级别数。计算熵时通常以图像的灰度直方图为基础，通过统计每个灰度级别的出现频率来估算概率。

Qabf（Quality Assessment via Blind Fusion）是一种新颖的客观非参考的图像质量评价指标，该指标利用局部度量来评估显著信息在融合图像中的表现。对于每个像素值，Qabf 都会计算一个局部质量度量，然后根据像素的权重进行加权平均，得到整个图像的质量度量。其具体的计算公式如式 (23) 所示：

$$Q(a,b,f) = \frac{1}{|W|}\sum_{\omega\in W}(\lambda(\omega)Q_0(a,f|\omega) + (1-\lambda(\omega))Q_0(b,f|\omega)) \tag{23}$$

式中，a和b是输入图像；f是融合后图像；W是图像的像素集合；$\lambda(\omega)$是像素ω的权重；$Q_0(a,f|\omega)$是像素ω在图像a和f之间的局部质量度量。

结构相似性（SSIM）是从模拟人类主观感知的角度出发，以客观指标来衡量融合图像与源图像之间的边缘和纹理的相似性。SSIM 从亮度、对比度以及结构 3 个方面量化图像的属性，使用均值来估计亮度，使用方差来估计对比度，使用协方差来估计结构相似程度。其计算方式如式 (24) 所示：

$$\mathrm{SSIM}(x,y) = [l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma} \tag{24}$$

式中，l表示亮度部分；c表示对比度部分；s表示结构部分；α，β 和 γ 用来调节三者在评估中的相对重要性。$l(x,y) = \dfrac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$；$c(x,y) = \dfrac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$；$s(\mathrm{x},\mathrm{y}) = \dfrac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$。$\mu$项为相应图像的均值，$\sigma$为相应图像的标准差，$C$项为常数。

SSIM 可以很好地反映融合后的图像在结构方面的保持程度。SSIM 值的范围为 0～1，越大代表图像的结构相似程度越高，如果融合图像能够很好地保留源图像的边缘和纹理信息，那么它的 SSIM 值就会接近 1；反之，如果融合图像与源图像在结构上有很大的差异，那么它的 SSIM 值就会接近 0。

3.5　实验结果与分析

3.5.1　CT 与 MRI 的融合结果分析

在临床医学应用中，CT 和 MRI 各自有其独特的优势。CT 能提供组织的密度信息，对骨骼等硬组织的显示效果好，而 MRI 对软组织的分辨率高，能显示出 CT 难以检测到的某些疾病。通过融合这两种图像，可以提高诊断的准确性，更准确地判断肿瘤的位置和大小，在放射治疗等治疗方法中可以提供更准确的病灶定位，提高治疗的精确性和有效性。

本节使用不同的多模态图像融合方法，在 CT 和其相应的 MRI 图像上进行了融合实验，通过 SD、PSNR、MI、VIF、EN、Qabf、SSIM 这 7 种客观指标对实验结果进行了分析，对比方法包括 Dif-fusion、MMIF-DDFM、MSRPAN、SDDGAN、U2Fusion，实验结果如表 1 所示。

表 1　CT 与 MRI 图像融合的客观量化指标

方法	SD↑	PSNR↑	MI↑	VIF↑	EN↑	Qabf↑	SSIM↑
Dif-Fusion	**87.01**	61.26	2.74	**0.49**	5.27	**0.56**	0.21
MMIF-DDFM	63.60	**63.79**	3.35	**0.47**	4.07	0.40	**0.68**
MSRPAN	**87.05**	61.15	**3.86**	0.44	3.82	0.43	**0.67**
SDDGAN	64.42	61.89	2.37	0.38	**4.97**	0.24	0.17
U2Fusion	57.23	**63.18**	2.41	0.38	4.56	0.47	0.21
本方法	84.37	62.85	**3.82**	**0.47**	**5.52**	**0.52**	**0.67**

注：加粗项为各指标前二的得分

根据表 1 中所示的实验结果可以发现，本文的方法在 CT 和 MRI 模态上生成的融合图像质量在各客观量化指标上都取得了不错的结果。在标准差（SD）方面，Dif-Fusion 和 MSRPAN 达到了 87.01 和 87.05 的高分，本文方法取得了 84.37 的分数，

分差在−3 以内；在熵（EN）方面，本文方法得到了最高的分数。在以上两个指标和其他几种方法拉开了较大差距，表明融合后的图像能包含更丰富的信息、细节和特征。峰值信噪比（PSNR）方面，各方法都能达到 60 dB 以上，都保持很高的保真度，且本文方法在该指标上保持前三水平。在互信息（MI）方面，本文方法取得了 3.82 的分数，与表现较好的 MSRPAN 的 3.86 分数分差在−0.04，水平相当，并和其他几种方法拉开了较大差距，表明融合后的图像与源图像相关性较高。在视觉信息保真度（VIF）方面，本文方法取得了 0.47 的分数，与最优的 Dif-fusion 方法差距在−0.03 之内，客观上来说融合后的图像相较于其他方法能更好地保留源图像的信息。在 Qabf 指标方面，本文方法取得了 0.52 的分数，稍逊于 Dif-Fusion。在结构相似性指数（SSIM）方面，本文方法和 DDFM、MSRPAN 取得了较高并接近的分数，客观上来说融合图像会更符合人类主观感知，且在边缘、纹理和局部质量上能保持和源图像一致的水平。

总体上，在 CT 和 MRI 融合场景下，Dif-fusion 在图像信息丰富程度和视觉保真度方面取得了较好的效果，MMIF-DDFM 在视觉保真度以及信息关联度、结构还原度方面效果较好。综合考虑各指标，其他方法虽然在某个别指标上能达到最优，但相比本文的方法提升微弱，且本文方法在所有的指标上均表现优异，均能达到前三或者最优，综合考虑各客观指标，本文方法展现出了一定的优势。

各方法在 CT 与 MRI 上的部分融合结果的可视化结果如图 5 所示。在前两行图像中，MRI 对于脑组织等软组织部分有较高的分辨率，CT 图像对于颅骨等高密度区域呈现为白色。MMIF-DDFM、SDDGAN、U2Fusion 的融合结果对于颅骨、脑实质中的高密度区域的呈现效果不佳。MMIF-DDFM 方法将图像转换为 YCrCb 空间后仅针对 Y 通道做融合，导致图像整体偏灰，且原本的高亮和灰度区域分界不明显。SDDGAN 强调对于源图像语义信息的理解，使用语义分割提取两种模态中的有效信息，根据源图像生成不同权重来融合多模态图像，该方法受网络对于图像的语义理解影响较大，语义理解混乱导致软组织处的处理不佳，细节模糊，失真较大。U2Fusion 对于结构边缘以及纹理细节的处理较好，但对于脑实质结构内部内容的融合不到位。相比之下，Dif-fusion、MSRPAN 和本文方法的融合结果能很好地呈现并区分出软组织和高密度区域。根据图中红色区域可观察出，本方法融合出的脑组织细节更加丰富。

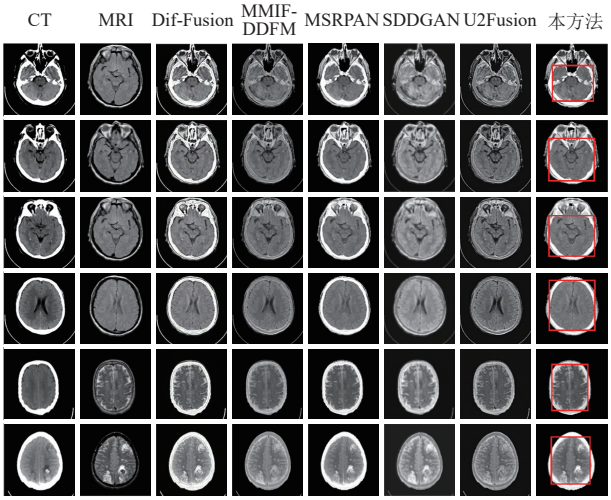


图 5 CT 和 MRI 融合可视化结果

综合主客观情况分析，当要进行 MRI 和 CT 图像融合时，本文方法具有一定的优势。在实际情况中，并非客观量化指标越高，融合结果的质量越好，仍需要根据不同的需求，选取合适的融合方法。

3.5.2 PET 与 MRI 的融合结果分析

PET 是正电子发射计算机断层显像分子成像，可以和 MRI 核磁共振成像中显著的软组织成像结合起来，实现对组织中扩散的疾病细胞成像的功能，且 PET 和 MRI 检查可在一次扫描中同时完成，因此 PET 和 MRI 成像有其独特的优势。在临床医学应用中，PET 和 MRI 图像融合常用于恶性肿瘤、神经系统和心血管疾病三大领域，是诊断和指导治疗各种恶性肿瘤、冠心病和脑部疾病的重要方法。

本节通过在 PET 和其相应的 MRI 图像上进行多模态图像融合实验，采用 SD、PSNR、MI、VIF、EN、Qabf、SSIM 7 种客观指标对实验结果进行客观分析，对比方法包括 Dif-Fusion、MMIF-DDFM、MSRPAN、SDDGAN、U2Fusion 方法，实验结果如表 2 所示。

表 2 PET 与 MRI 图像融合的客观量化指标

方法	SD↑	PSNR↑	MI↑	VIF↑	EN↑	Qabf↑	SSIM↑
Dif-Fusion	82.05	61.22	2.61	0.52	**6.18**	0.59	0.28
MMIF-DDFM	71.85	**63.44**	3.92	**0.68**	5.12	0.58	0.58
MSRPAN	**83.48**	60.96	**4.09**	**0.62**	4.92	**0.65**	**0.62**
SDDGAN	76.20	62.00	2.41	0.42	6.01	0.30	0.23
U2Fusion	43.18	**62.40**	2.51	0.47	5.07	0.25	0.20
本方法	**82.97**	62.23	**4.11**	0.59	**6.39**	**0.65**	**0.61**

注：加粗项为各指标前二的得分

实验结果表明，在 PET 和 MRI 融合场景下，

Dif-fusion 方法更关注于融合图像的信息丰富程度，MMIF-DDFM 更关注于融合图像的结构保持程度，MSRPAN 在多个指标达到了最优或次优，客观上表现优异，SDDGAN、U2Fusion 在本场景未表现出明显优势。在标准差 SD 方面，本文方法与 MSRPAN 分差在−1 分以内。在熵 EN 方面，本文方法取得了最高的分数，客观表明生成的融合图像更为复杂且信息丰富。在 Qabf 和互信息（MI）方面，本文方法也达到了最优，客观上表明融合后的图像与源图像有更高的关联性，同时有更好的视觉质量。在视觉保真度 VIF 和峰值信噪比 PSNR 方面，本文方法表现良好，稍逊于 MMIF-DDFM，客观上融合后的图像能较好地保留源图像的信息，同时 PSNR 保持在 60 dB 以上表明融合图像有很好的保真度。在结构相似度 SSIM 方面，MSRPAN 和本文方法效果上领先其他方法，客观上展现出较好的结构保持性能且更符合人眼视觉观感。

各方法在 PET 与 MRI 上的部分融合结果的可视化结果如图 6 所示。

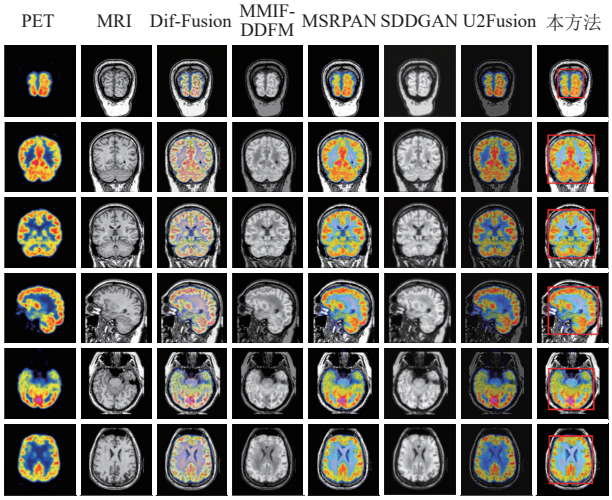


图 6　PET 和 MRI 融合可视化结果

第一行 PET 图像中红色区域表示大量的示踪剂积聚，表明该区域是活跃组织或是病变区域，如肿瘤、炎症或高代谢的正常组织。第二行 MRI 图像中对于脑组织的分辨率较高，展现了更多的纹理细节特征。MMIF-DDFM 由于只对 YCRCb 的 Y 通道做了融合，融合后的图像是一张灰度图，不带颜色信息，SDDGAN 融合后也只输出了单通道灰度图，二者在 PET 图像高亮的区域更多关注了亮度的变化，而缺少了色彩信息，虽然在客观量化指标表现尚可，但主观视角下对于活跃组织的呈现效果不佳。同时，在脑实质区域，亮度的变化干扰了细节特征的还原。U2Fusion 对于 MRI 中高信号强度区域还原不佳，整体偏灰，亮度较暗，视觉质量较低，但对于被 PET 中红色区域覆盖的软组织的细节保留较好。Dif-fusion 对于 MRI 中的细节纹理特征保留较好，但是融合后属于 PET 的示踪剂积聚区的色彩失真较大。相比之下，本文方法和 MSRPAN 都呈现了很好的视觉观感和色彩还原度，其中，MSRPAN 的结果中示踪剂积聚区掩盖了脑实质的纹理细节特征。根据图中红色区域可观察出，本文方法在保留了示踪剂积聚区的视觉保真度外，更强化了脑组织纹理细节的表达。

综合主客观情况分析，本文方法在各客观指标上均达到了最优或是和最优方法保持微弱的差距，同时可视化结果也展示了很好的结构保持性和视觉保真度，在 PET 和 MRI 融合场景下表现出比较均衡全面的融合性能。在实际情况中，可以根据不同的需求和关注重点，选取合适的图像融合方法。

3.5.3　SPECT 与 MRI 的融合结果分析

SPECT 是单光子发射计算机断层成像术，在代谢活跃的区域能提供较好的显示效果，但在软组织区域显示效果不佳。而 MRI 在软组织区域有较高的分辨率，与 SPECT 图像融合有利于提供更全面、准确的医疗图像信息，提高诊断的准确性，如帮助医生更准确地判断肿瘤的位置和大小、提供更精确的病灶定位。

本节通过在 SPECT 图像和其相应的 MRI 图像上进行多模态图像融合实验，对比方法包括 Dif-Fusion、MMIF-DDFM、MSRPAN、SDDGAN、U2Fusion 这 5 种先进的方法，并采用 SD、PSNR、MI、VIF、EN、Qabf、SSIM 共 7 种客观指标对实验结果进行了客观量化分析，实验结果如表 3 所示。

表 3　SPECT 与 MRI 图像融合的客观量化指标

方法	SD↑	PSNR↑	MI↑	VIF↑	EN↑	Qabf↑	SSIM↑
Dif-Fusion	59.62	66.29	2.82	**0.56**	**5.34**	**0.61**	0.25
MMIF-DDFM	60.33	66.93	3.55	0.55	4.55	0.59	**0.61**
MSRPAN	54.91	67.37	**4.02**	0.54	4.34	0.56	**0.61**
SDDGAN	**67.95**	63.25	2.65	0.43	5.35	0.34	0.20
U2Fusion	32.36	**67.69**	2.50	0.41	4.32	0.34	0.20
本方法	**73.00**	**67.59**	**4.06**	**0.56**	**5.62**	**0.67**	**0.61**

注：加粗项为各指标前二的得分

根据表中所示的实验结果可以发现，本文方法在 SPECT 和 MRI 模态上生成的融合图像质量在各客观量化指标上都取得了显著的优势。在标准差

（SD）方面本文方法取得了 73.00 的分数，在互信息（MI）方面本文方法取得了 4.06 的分数，均明显领先于其他方法，表明融合图像具有更好的细节和特征，且较好地保留了源图像信息。在峰值信噪比（PSNR）方面，本文方法取得了 67.59 的分数，位列前三水平，说明融合图像在保真度上表现较好。在视觉信息保真度（VIF）方面，本文方法取得了 0.56 的分数，客观上来说融合后的图像相较于其他方法能更好地保留源图像的信息。在 Qabf 指标方面，本文方法取得了 0.67 的分数，较其他方法明显更好，显示融合图像在结构相似性方面更符合人类主观感知。在结构相似性指数（SSIM）方面，本文方法与 DDFM、MSRPAN 取得了较高且接近的分数，表明融合图像在边缘、纹理和局部质量上能保持和源图像一致的水平。

各方法在 SPECT 与 MRI 上的部分融合结果的可视化结果如图 7 所示。第一行 SPECT 图像中红色和紫色区域表示了放射性示踪剂的分布，这些区域代表可能是高代谢区域或是病变区域。第二行 MRI 图像中对于脑组织的分辨率较高，展现了更多的纹理细节特征。MMIF-DDFM 和 SDDGAN 方法的融合结果均保留了单通道的融合图像，丢失了色彩信息，色彩强度的变化通过亮度变化的形式呈现。其中 MMIF-DDFM 对于能较好地还原 MRI 中的细节纹理特征，SDDGAN 的融合结果存在大量的噪声。U2Fusion 的结果对于 MRI 中的结构细节还原不佳，整体偏灰暗，视觉质量较低，但对于 SPECT 图像和 MRI 图像重叠区域的细节保留比较好。Dif-fusion 对于 MRI 中的细节纹理特征保留较好，但色彩失真较大。MSRPAN 的色彩还原程度比较好，视觉保真度高，但在示踪剂覆盖区域下的脑组织的纹理细节特征有部分被红色区域掩盖。相比之下，本文方法和 Dif-fusion 对于脑组织细节纹理特征的保留性会更好，和 SPECT 图像中示踪剂积聚区域的融合效果比较好。其中，Dif-fusion 在色彩上呈现了一定的失真，本文方法在同时兼顾了色彩的还原以及脑组织纹理细节的表达。

综合而言，本文方法在 SPECT 与 MRI 图像上的融合结果在多个客观指标上都表现出色，表现出较好的图像质量和信息丰富度。除此之外，MMIF-DDFM 方法在多个指标上也达到了最优或次优的表现，客观上表明在 SPECT 和 MRI 图像融合任务上也有较好的表现。在可视化结果的主观分析上，MMIF-DDFM 虽然较好地保留了结构纹理细节特征，但失去了丰富的色彩表达。根据图中红色区域可观察出，本文方法在保留 MRI 图像中的细节信息的同时能兼顾到 SPECT 图像中色彩的还原。在实际情况中，可以根据不同的需求和关注重点，选取合适的图像融合方法。

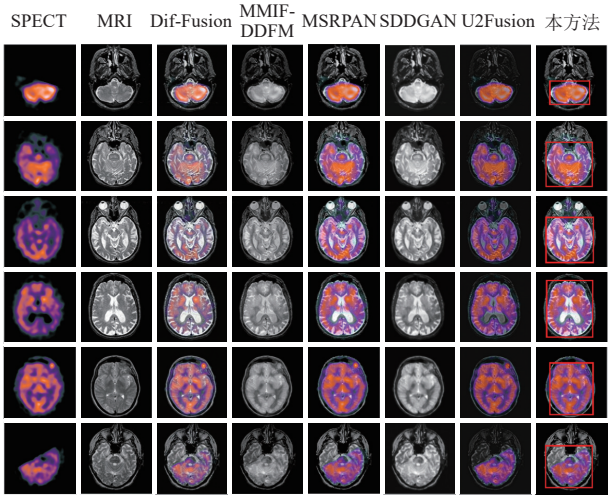


图 7 SPECT 和 MRI 融合可视化结果

3.5.4 消融实验

为了验证扩散过程在提取多通道互补特征、提高色彩保真度和视觉质量上的有效性，本文设计了消融实验。具体来说，在保持去噪网络结构不变的情况下，取消了扩散过程以及损失函数中的扩散损失项，为了保持输入通道的一致性，将去噪网络的输入中去噪图像 1 所在通道图像和去噪图像 2 所在通道图像替换为未受噪声干扰的源图像。消融实验的实验结果如表 4 所示。

表 4 扩散过程消融实验

模态	方法	SD↑	PSNR↑	MI↑	VIF↑	EN↑	Qabf↑	SSIM↑
CT-	w/diff.	84.37	62.85	3.82	0.47	5.52	0.52	0.67
MRI	wo/diff.	85.52	60.78	3.42	0.43	4.94	0.36	0.67
PET-	w/diff.	82.97	62.23	4.11	0.59	6.39	0.65	0.61
MRI	wo/diff.	75.45	60.34	4.71	0.55	5.80	0.49	0.58
SPECT-	w/diff.	73.00	67.59	4.06	0.56	5.62	0.67	0.61
MRI	wo/diff.	64.32	65.94	3.69	0.50	5.05	0.52	0.59

本实验比较了带扩散过程与不带扩散过程的多模态图像融合方法在 CT-MRI、PET-MRI 以及 SPECT-MRI 等不同医学成像模态上的性能差异。由实验结果可知，在 AANLIB 数据集中，去除扩散过程后，本文方法在多个指标上出现了不同程度的性能波动。根据实验结果，观察到在不采用扩散过程的情况下，在 CT 和 MRI 模态融合的场景中

获得了较高的标准差值，这可能是在融合单通道灰度图像时去除扩散过程后引入了更多其他通道的不相干噪声导致，同时在其他评价指标上均体现出一定的劣势；但在结构相似度指标上，在结构相似度损失的帮助下，去除扩散过程前后并未拉开太大差距。值得注意的是，在 PSNR、MI、VIF 及 EN 等反映信息保持和图像内容丰富程度的指标上，带扩散过程的方法往往表现更优，同时在 Qabf 这一特定评估维度也取得了一定的优势，反映出引入扩散过程对于提升融合结果的视觉质量是一个比较有效的选择。

为了验证判别器在引导去噪网络生成合适噪声的有效性，本文设计了消融实验。具体来说，在保持其他模块不变的情况下，取消了双判别器及相关损失。实验结果如表 5 所示。

表 5　判别器消融实验

模态	方法	SD↑	PSNR↑	MI↑	VIF↑	EN↑	Qabf↑
CT-MRI	w/discrim.	84.37	62.85	3.82	0.47	5.52	0.52
	wo/ discrim.	80.00	61.32	3.20	0.52	4.71	0.61
PET-MRI	w/discrim.	82.97	62.23	4.11	0.59	6.39	0.65
	wo/ discrim	84.00	60.75	3.09	0.57	5.95	0.65
SPECT-MRI	w/discrim.	73.00	67.59	4.06	0.56	5.62	0.67
	wo/ discrim	59.50	65.87	3.34	0.67	5.18	0.67

通过以上在 3 个多模态数据的实验可看出，双判别器在多个指标中都有效提升了融合的效果。

4　结束语

本文介绍了一种基于去噪扩散模型的多模态图像融合方法，通过将推理融合解决方案集成到扩散迭代去噪过程中，克服单一端到端网络学习能力有限的问题。对扩散去噪过程针对多模态医学图像融合任务做出了优化，引入了两个模态判别器加强去噪网络对于模态特征的理解，利用先验和源图像的跨模态信息生成高质量的融合图像。通过 AANLIB 数据集上 3 种模态下的融合实验，探讨对比了不同方法的融合质量，并对融合结果进行了客观量化指标的分析，以及更符合人眼主观视觉的可视化分析。未来工作中可以向探索动态时间步调度策略优化以及三维体数据融合扩展。

参 考 文 献

[1] ABBASI S, TAVAKOLI M, BOVEIRI H R, et al. Medical image registration using unsupervised deep neural network: A scoping literature review[J]. Biomedical Signal Processing and Control, 2022, 73: 103444.

[2] 刘思捷. 智慧医疗: 医学图像融合领域技术综述[J]. 专利代理, 2023(1): 39-45.
LIU S J. Intelligent medical treatment: A review of medical image fusion technology[J]. Patent Agency, 2023(1): 39-45.

[3] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[4] 石争浩, 仵晨伟, 李成建, 等. 航空遥感图像深度学习目标检测技术研究进展[J]. 中国图象图形学报, 2023, 28(9): 2616-2643.
SHI Z H, WU C W, LI C J, et al. Object detection techniques based on deep learning for aerial remote sensing images: A survey[J]. Journal of Image and Graphics, 2023, 28(9): 2616-2643.

[5] 罗福亚. 基于视觉认知机理的夜间热红外图像彩色化方法研究[D]. 成都: 电子科技大学, 2023.
LUO F Y. Research on colorization method of nighttime thermal infrared images based on visual cognitive mechanism[D]. Chengdu: University of Electronic Science and Technology of China, 2023.

[6] 王丽芳, 米嘉, 秦品乐, 等. 改进 U-Net3+与跨模态注意力块的医学图像融合[J]. 中国图象图形学报, 2022, 27(12): 3622-3636.
WANG L F, MI J, QIN P L, et al. Medical image fusion using improved U-Net3+ and cross-modal attention blocks[J]. Journal of Image and Graphics, 2022, 27(12): 3622-3636.

[7] HERMESSI H, MOURALI O, ZAGROUBA E. Multimodal medical image fusion review: Theoretical background and recent advances[J]. Signal Processing, 2021, 183: 108036.

[8] YANG B, JING Z L, ZHAO H T. Review of pixel-level image fusion[J]. Journal of Shanghai Jiaotong University (Science), 2010, 15(1): 6-12.

[9] DU J, LI W S, LU K, et al. An overview of multi-modal medical image fusion[J]. Neurocomputing, 2016, 215: 3-20.

[10] SINGH S, ANAND R S. Ripplet domain fusion approach for CT and MR medical image information[J]. Biomedical Signal Processing and Control, 2018, 46: 281-292.

[11] SHABANZADE F, GHASSEMIAN H. Combination of wavelet and contourlet transforms for PET and MRI image fusion[C]//Proceedings of the Artificial Intelligence and Signal Processing Conference. New York: IEEE, 2017: 178-183.

[12] JIN Z Y, WANG Y J, CHEN Z G, et al. Medical image fusion in gradient domain with structure tensor[J]. Journal of Medical Imaging and Health Informatics, 2016, 6(5): 1314-1318.

[13] BHAVANA V, KRISHNAPPA H K. A survey on multi-modality medical image fusion[C]//Proceedings of the International Conference on Wireless Communications, Signal Processing and Networking. New York: IEEE, 2016: 1326-1329.

[14] EGFIN NIRMALA D, VAIDEHI V. Comparison of Pixel-level and feature level image fusion methods[C]//

Proceedings of the 2nd International Conference on Computing for Sustainable Global Development. New York: IEEE, 2015: 743-748.

[15] XI X X, LUO X Q, ZHANG Z C, et al. Multimodal medical volumetric image fusion based on multi-feature in 3-D shearlet transform[C]//Proceedings of the International Smart Cities Conference. New York: IEEE, 2017: 1-6.

[16] TALBAR S N, CHAVAN S S, PAWAR A. Non-subsampled complex wavelet transform based medical image fusion[C]//Proceedings of the Future Technologies Conference (FTC) 2018. Cham: Springer International Publishing, 2018: 548-556.

[17] TANG L, LI L D, QIAN J S, et al. NSCT-based multimodal medical image fusion with sparse representation and pulse coupled neural network[J]. Journal of Information Hiding and Multimedia Signal Processing, 2016, 7(6): 1306-1316.

[18] MA J Y, MA Y, LI C. Infrared and visible image fusion methods and applications: A survey[J]. Information Fusion, 2019, 45: 153-178.

[19] VELMURUGAN S P, SIVAKUMAR P, RAJASEKARAN M P. Multimodality image fusion using centre-based genetic algorithm and fuzzy logic[J]. International Journal of Biomedical Engineering and Technology, 2018, 28 (4): 322.

[20] SHAHDOOSTI H R, MEHRABI A. Multimodal image fusion using sparse representation classification in tetrolet domain[J]. Digital Signal Processing, 2018, 79: 9-22.

[21] 黄渝萍, 李伟生. 医学图像融合方法综述[J]. 中国图象图形学报, 2023, 28(1): 118-143.
HUANG Y P, LI W S. A review of medical image fusion methods[J]. Journal of Image and Graphics, 2023, 28(1): 118-143.

[22] SONG J M, MENG C L, ERMON S. Denoising diffusion implicit models[EB/OL]. [2024-09-19]. https://arXiv.org/abs/2010.02502.

[23] ZHOU H B, WU W, ZHANG Y D, et al. Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network[J]. IEEE Transactions on Multimedia, 2021, 25: 635-648.

[24] XU H, MA J Y, JIANG J J, et al. U2Fusion: A unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 502-518.

[25] FU J, LI W S, DU J, et al. A multiscale residual pyramid attention network for medical image fusion[J]. Biomedical Signal Processing and Control, 2021, 66: 102488.

[26] YUE J, FANG L Y, XIA S B, et al. Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models[J]. IEEE Transactions on Image Processing, 2023, 32: 5705-5720.

责任编辑 张 莉