软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(5):2134−2151 [doi: 10.13328/j.cnki.jos.006772]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

预训练模型特征提取的双对抗磁共振图像融合网络研究*

刘 慧[1,2], 李珊珊[1,2], 高珊珊[1,2], 邓 凯[3], 徐 岗[4], 张彩明[2,5]

[1](山东财经大学 计算机科学与技术学院, 山东 济南 250014)

[2](山东省数字媒体技术重点实验室, 山东 济南 250014)

[3](山东第一医科大学第一附属医院, 山东 济南 250013)

[4](杭州电子科技大学 计算机学院, 浙江 杭州 310018)

[5](山东大学 软件学院, 山东 济南 250101)

通信作者: 张彩明, E-mail: czhang@sdu.edu.cn

摘 要: 随着多模态医学图像在临床诊疗工作中的普及, 建立在时空相关性特性基础上的融合技术得到快速发展, 融合后的医学图像不仅可以保留各模态源图像的独有特征, 而且能够强化互补信息、便于医生阅片. 目前大多数方法采用人工定义约束的策略来实现特征提取和特征融合, 这容易导致融合图像中部分有用信息丢失和细节不清晰等问题. 为此, 提出一种基于预训练模型特征提取的双对抗融合网络实现 MR-T1/MR-T2 图像的融合. 该网络由一个特征提取模块、一个特征融合模块和两个鉴别网络模块组成. 由于已配准的多模态医学图像数据集规模较小, 无法对特征提取网络进行充分的训练, 又因预训练模型具有强大的数据表征能力, 故将预先训练的卷积神经网络模型嵌入到特征提取模块以生成特征图. 然后, 特征融合网络负责融合深度特征并输出融合图像. 两个鉴别网络通过对源图像与融合图像进行准确分类, 分别与特征融合网络建立对抗关系, 最终激励其学习出最优的融合参数. 实验结果证明了预训练技术在所提方法中的有效性, 同时与现有的 6 种典型融合方法相比, 所提方法融合结果在视觉效果和量化指标方面均取得最优表现.

关键词: 多模态医学图像; 图像融合; 预训练模型; 双鉴别网络; 对抗学习

中图法分类号: TP391

中文引用格式: 刘慧, 李珊珊, 高珊珊, 邓凯, 徐岗, 张彩明. 预训练模型特征提取的双对抗磁共振图像融合网络研究. 软件学报, 2023, 34(5): 2134–2151. http://www.jos.org.cn/1000-9825/6772.htm

英文引用格式: Liu H, Li SS, Gao SS, Deng K, Xu G, Zhang CM. Research on Dual-adversarial MR Image Fusion Network Using Pre-trained Model for Feature Extraction. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2134–2151 (in Chinese). http://www.jos.org.cn/1000-9825/6772.htm

Research on Dual-adversarial MR Image Fusion Network Using Pre-trained Model for Feature Extraction

LIU Hui[1,2], LI Shan-Shan[1,2], GAO Shan-Shan[1,2], DENG Kai[3], XU Gang[4], ZHANG Cai-Ming[2,5]

[1](School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China)

[2](Shandong Key Laboratory of Digital Media Technology, Jinan 250014, China)

[3](The First Affiliated Hospital of Shandong First Medical University, Jinan 250013, China)

[4](College of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China)

[5](School of Software, Shandong University, Jinan 250101, China)

* 基金项目: 国家自然科学基金 (62072274, U1909210); 山东省科技成果转移转化项目 (2021LYXZ011); 浙江省重点研发计划 (2021C01108)
本文由“融合预训练技术的多模态学习研究”专题特约编辑宋雪萌副教授、聂礼强教授、申恒涛教授、田奇教授、黄华教授推荐.
收稿时间: 2022-04-18; 修改时间: 2022-05-29; 采用时间: 2022-08-24; jos 在线出版时间: 2022-09-20
CNKI 网络首发时间: 2023-03-17

Abstract: With the popularization of multimodal medical images in clinical diagnosis and treatment, fusion technology based on spatiotemporal correlation characteristics has been developed rapidly. The fused medical images not only retain the unique features of source images with various modalities but also strengthen the complementary information, which can facilitate image reading. At present, most methods perform feature extraction and feature fusion by manually defining constraints, which can easily lead to the loss of useful information and unclear details in the fused images. In light of this, a dual-adversarial fusion network using a pre-trained model for feature extraction is proposed in this study to fuse MR-T1/MR-T2 images. The network consists of a feature extraction module, a feature fusion module, and two discriminator network modules. Due to the small scale of the registered multimodal medical image dataset, the feature extraction network cannot be fully trained. In addition, as the pre-trained model has powerful data representation ability, a pre-trained convolutional neural network model is embedded into the feature extraction module to generate the feature map. Then, the feature fusion network fuses the deep features and outputs fused images. Through accurate classification of the source and fused images, the two discriminator networks establish adversarial relations with the feature fusion network separately and eventually encourage it to learn the optimal fusion parameters. The experimental results illustrate the effectiveness of pre-trained technology in this method. Compared with six existing typical fusion methods, the proposed method can generate the fused results of optimal performance in visual effects and quantitative metrics.

Key words: multi-modal medical image; image fusion; pre-trained model; dual-adversarial network; adversarial learning

随着医学成像技术的快速发展, 多模态医学图像在疾病诊断、治疗、手术规划等多种临床应用中发挥着越来越重要的作用[1]. 由不同成像设备获取的不同模态医学图像可以向医生提供更加全面的临床信息. 例如, 电子计算机断层扫描 (computed tomography, CT) 图像主要反映骨骼或植入物等致密结构信息; 正电子发射型计算机断层扫描 (positron emission computed tomography, PET) 和单光电子发射型计算机断层扫描 (single photon emission computed tomography, SPECT) 图像侧重于提供血流和代谢变化等功能性信息. 磁共振 (magnetic resonance, MR) 根据弛豫时间的不同可分为 MR-T1 和 MR-T2. 其中 MR-T1 图像主要反映器官或组织的解剖信息; MR-T2 图像对出血比较敏感, 有利于准确观察病变组织. 然而, 每种成像方式都有其自身的特点和应用局限性, 需要开发可直接获取混合图像的成像设备, 或者探索可融合现有不同模态医学图像的有效技术. 与前者相比, 开发有效的图像融合方法成本更低、时间更短. 因多模态医学图像融合方法能够将不同模态医学图像中的关键特征和互补信息进行融合, 并将融合图像可视化, 帮助医生更加直观、科学地进行各种目的的决策, 所以近年来受到研究者们的广泛关注. 除此之外, 多模态医学图像融合还能够为医学图像分割[2]以及多视图聚类[3]等领域提供理论支持和参考, 具有重要的研究价值. 不同模态医学图像特征分布多样, 研究和分析方法也不尽相同. 本文主要针对医学图像中的 MR-T1/MR-T2 图像进行分析并实现融合. 图 1 中, 前两张图像为待融合的源图像, 第 3 张图像是前两张图像的融合结果. 由图 1 及局部放大图像可以看出, 融合后的医学图像不仅需要充分保留源图像的结构特征, 而且需要将不同模态的显著特征和互补信息进行有效融合, 以此才能达到辅助医生快速定位病灶并诊断疾病的目的. 由于不同模态医学图像的特征是复杂多样的, 因此如何充分提取并融合医学图像的各类特征成为当前解决该问题的主要难点[4,5].

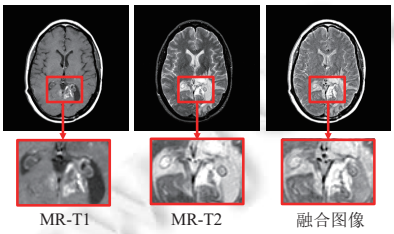


图 1 MR-T1/MR-T2 图像及其融合图像

目前, 国内外学者提出了多种用于解决图像融合问题的有效方法. 其中, 基于深度学习的融合方法可通过神经网络层的堆叠, 自适应地实现特征提取、特征融合以及图像重构, 故被广泛用于各种图像融合任务[6]. 例如, Prabhakar 等人[7]提出了一种基于卷积神经网络的深度学习架构, 有效地实现了静态多曝光图像的融合. Li 等人[8]提出了一种用于红外与可见光图像融合的深度学习方法, 并通过有效结合卷积层和密集块以获取更多有用特征.

然而, 此类图像融合方法通常依赖于卷积神经网络 (convolutional neural networks, CNN) 模型, 故需要经过大量有标签数据的训练后才能获得强大的数据表征能力. 并且在训练样本较少的情况下, 深度神经网络容易出现过拟合问题, 在实际应用中泛化能力有限. 为解决这一问题, 预训练技术应运而生. 研究表明[9], 由于复杂的网络结构和庞大的模型参数, 预训练技术能够有效地从大量已标记和未标记的数据中捕获先验知识, 并将其存储到模型参数中. 在解决特定任务的过程中, 通过对预训练模型进行微调, 能够将隐含在参数中的丰富先验知识应用于各种下游任务. 因此, 在计算机视觉领域, 面对小数据集的图像处理任务时, 很多研究选择采用预训练模型提取图像特征, 然后再完成相应的下游任务. 如 Iglovikov 等人[10]使用预先训练的编码器对 U-Net 进行改进, 有效地实现了图像分割, 并展示了 U-Net 的性能是如何通过使用预先训练的权值得到提高的. 此外, 预训练技术也被广泛应用于处理多模态问题. 例如 Chen 等人[11]利用预训练学习视觉和语言任务中的通用图像-文本表示, 出色地完成了视觉问答、图像-文本检索、视觉常识推理等多个视觉和语言任务. 而在多模态医学融合任务当中, 已完成配准的多模态医学图像数据集规模一般较小, 不足以训练出具有强大特征表示和特征提取能力的网络模型. 同时, 多模态医学图像融合属于非监督任务, 大多数基于深度学习的融合方法需依据经验设计复杂的损失函数来约束和引导融合模型的优化, 这不仅容易导致图像纹理细节等结构特征的丢失, 而且增大了网络的训练难度.

针对上述问题, 本文提出一种基于预训练模型特征提取的双对抗融合网络. 该网络由特征提取模块、特征融合模块和双对抗网络模块 3 部分组成. 其中特征提取模块集成了在 ImageNet 数据集[12]上预训练的 CNN 模型, 通过将该预训练模型的原有参数作为初始化数据, 然后在网络训练过程中进一步对参数进行微调以提取融合特征. 此外, 本文定义了基于 DenseNet[13]的融合网络来实现特征融合以及融合图像输出. 同时, 针对不同模态图像的特征, 设计了对称鉴别网络分别与融合网络建立对抗关系, 在不依赖复杂损失函数的情况下, 自适应地约束和引导网络进行参数更新.

本文的主要贡献如下.

(1) 将预训练的特征提取模型集成到多模态医学图像融合任务中, 解决了因配准的多模态医学图像样本数量少而导致的特征提取模型过拟合问题, 有效降低了网络的训练复杂度.

(2) 定义双鉴别网络分别与特征融合网络建立对抗关系, 通过对抗训练激励网络进行参数更新. 从而避免了人工设计过于复杂的损失函数, 并充分保留了各种不同模态的特征, 有助于提高融合性能.

(3) 将本文方法与 6 种现有的图像融合方法进行了实验对比, 通过可视化融合结果以及计算多种常用的融合图像评价指标证明了本文方法的有效性.

本文第 1 节介绍多模态医学图像融合的相关工作和研究现状. 第 2 节详细介绍基于预训练模型特征提取的双对抗融合网络. 第 3 节通过对比实验验证所提模型的有效性. 第 4 节总结全文, 并对未来研究做出展望.

1 相关工作

1.1 多模态医学图像融合

多模态医学图像融合是多源图像融合领域中一个极具特色的研究分支, 它的发展建立在多源图像融合研究的基础上, 同时与医学成像技术的发展密不可分. 早期的图像融合技术大多基于传统数学方法进行特征提取、特征融合以及图像重构. 经典的图像融合算法主要包括以下几类: 基于多尺度变换的方法[14] (例如: 基于非下采样轮廓波变换的方法[15])、基于稀疏表示的方法[16]、基于子空间的方法[17]和基于显著性的方法[18]. 然而, 上述方法大都需要根据经验人工设计图像分解、重构方案以及融合规则[19]. 由于不同模态的图像拥有不同的数据分布特点, 仅依靠人工设计的特征提取和融合规则很难将图像的全部特征进行充分表达和融合, 这将会导致某些纹理细节甚至关键信息的丢失.

后来, 由于深度学习在图像表示、特征提取以及融合规则的学习等方面显示出独特优势, 能够有效代替传统方法中的部分人工操作, 从而被广泛应用于图像融合领域. 例如, Yin 等人[20]通过非下采样剪切波变换 (nonsubsampled shearlet transform, NSST) 对源图像进行分解, 然后结合参数自适应的脉冲耦合神经网络 (parameter-

adaptive pulse-coupled neural network, PAPCNN) 实现了多模态医学图像融合. Zhang 等人[21]提出了一种基于梯度和强度比例保持 (proportional maintenance of gradient and intensity, PMGI) 的快速统一图像融合网络, 该网络通过分别提取并融合图像的梯度特征和像素强度特征, 进而实现端到端的图像融合. 但由于深度神经网络结构相对复杂, 且参数量巨大, 所以在处理图像融合这类训练数据规模较小的任务时, 通常会表现出过拟合, 泛化能力差等缺点.

为此, 不少研究通过将预训练模型引入到图像融合任务当中, 利用预训练模型强大的数据表征能力对融合特征进行合理地表示和充分地提取, 再根据不同的方式对特征进行融合. 例如, Lahoud 等人[22]提出了一种零学习快速医学图像融合 (zero-learning fast medical image fusion, ZLFMIF) 方法, 该方法采用在 ImageNet 数据集上预训练的 VGG 模型提取不同模态医学图像的特征, 并结合 Softmax 算子对特征进行融合以生成融合图像. 文献 [23] 提出了一种基于卷积神经网络的图像融合 (image fusion based on convolutional neural network, IFCNN) 模型. 该方法同样使用在 ImageNet 数据集上预训练的 CNN 模型提取图像特征, 但该方法考虑到预训练模型提取的特征本来是用于分类任务的, 故在预训练模型之后增加了一层卷积以调整特征使其适用于图像融合. 然而, 以上方法仅利用预训练模型原有的参数提取特征, 并未针对融合任务调整参数, 这会在很大程度上影响融合性能.

此外, 图像融合是一类缺乏 ground-truth 的非监督任务, 基于深度学习的融合方法普遍采用人工定义损失函数的方法对模型的训练过程进行约束和指导. 为了更加全面地保留图像特征, 损失函数被设计得越来越复杂, 这在很大程度上增加了网络训练的难度. 针对这一问题, Ma 等人[24]提出融合方法 FusionGAN, 通过引入生成对抗网络 (generative adversarial network, GAN)[25], 利用生成网络与鉴别网络之间的对抗学习来自动完成可见光与红外图像的特征提取及融合, 避免了人工设计复杂的损失函数. 但该方法中鉴别网络仅针对红外特征与生成网络形成单一对抗, 容易导致梯度特征的丢失, 不适用于医学图像融合任务.

鉴于此, 本文基于 GAN 框架提出了一种预训练模型特征提取的双对抗融合网络. 为充分提取特征, 该网络集成了在 ImageNet 大型数据集上预训练的 VGG-16 模型, 同时将模型参数进行微调以使其适用于多模态医学图像融合任务. 训练过程中, 定义两个鉴别网络分别针对不同模态的图像特征与融合网络建立双对抗关系, 以确保各类特征都能够得到充分保留.

1.2 预训练技术

随着深度学习的发展, 预训练技术在各研究领域均被广泛应用. 预训练的目的是从一个或多个源任务中, 经过大量数据的训练后获得足够的先验知识, 然后用于解决目标任务中的新问题. 其中源任务和目标任务可能具有不同的数据空间和处理目标, 但解决问题所需的先验知识必须一致[26]. 比如, 无论是图像分类还是图像融合, 都需要获取图像的深层特征并针对特征进行相应的计算和处理. 总的来说, 预训练技术可分为特征预训练[27]与参数预训练[28]两种. 特征预训练首先将跨领域或跨任务的知识进行预编码, 然后将其注入到目标任务中, 以实现知识迁移并提高目标任务的模型性能. 参数预训练能够将知识预编码为可共享的模型参数. 目标任务通过直接共享模型参数的先验分布, 或者对模型参数进行微调来提高模型表现.

近几年, 预训练技术在计算机视觉领域取得了较好的效果. 在完成图像处理任务的过程中, 多数研究一般采用一个基于 ImageNet 数据集的预先训练好的网络, 然后再进行微调以使其适用于下游任务[29]. 预训练技术能够明显改善模型的鲁棒性和不确定性的估计[30], 并且能够为下游任务提供更加丰富的特征. 基于上述成功应用, 本文拟探索预训练模型在多模态医学图像融合任务中的作用.

1.3 GAN 网络

近年来, GAN 在无监督图像处理领域发展迅速, 其特点是能够通过一种间接的方式对一个未知分布进行建模[31]. 相比于其他的图像生成模型, GAN 可利用网络内部的对抗训练自动完成模型优化, 不需要人为定义复杂的损失函数, 因此能够在不影响生成图像质量的同时降低计算复杂度. 多模态医学图像融合作为一类典型的无监督图像生成问题, 适合使用 GAN 来解决.

传统的 GAN 由一个生成网络和一个鉴别网络组成. 生成网络的目的是根据输入数据 z 生成所需图像 $G(z)$, 同时要保证生成图像看起来足够逼真. 而鉴别网络的目的是能够将生成图像和真实图像进行准确分类, 即将真实图

像判断为真, 将生成图像判断为假. 因此鉴别网络的输入是一张图像 x, 输出是一个概率 $D(x)\in[0,1]$. 其中 $D(x)$ 越接近 1, 说明 x 属于真实图像的概率越大. 在训练过程中, 生成网络需不断更新参数以生成更加逼真的图像去欺骗鉴别网络. 而鉴别网络也需根据生成网络的优化逐步提高自己的分类能力. 通过这种方式, 生成网络与鉴别网络动态地进行对抗博弈, 直到双方达到纳什均衡[32]. 此时鉴别网络不再能够对真实图像和生成图像进行准确分类, 模型收敛, 训练结束. 值得注意的是, 生成网络与鉴别网络是两个相对独立的网络结构. 为避免两者在参数更新时发生混乱, 生成网络与鉴别网络需交替训练. 当生成网络参数固定时, 鉴别网络根据公式 (1) 更新参数:

$$\max_{D} V(D,G)=\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)]+\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{1}$$

其中, $\mathbb{E}[\cdot]$表示数据的期望, p_{data} 与 p_z 分别表示真实数据与输入数据的分布. 当鉴别网络参数固定时, 生成网络根据公式 (2) 更新参数:

$$\min_{G} V(D,G)=\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{2}$$

尽管传统的 GAN 在图像生成领域已经取得很大成功, 但依然存在两个关键问题: 一是如何避免模式崩塌, 改善生成图像的质量; 二是如何平衡生成器和鉴别器的训练程度, 提高训练过程的稳定性. 为解决这两个问题, 许多研究者提出了针对 GAN 的改进方案. 例如: Radford 等人[33]提出 DCGANs (deep convolutional generative adversarial networks), 该方法通过对鉴别网络和生成网络的架构进行实验枚举, 最终找到一组比较好的网络架构设置, 但这只能在一定程度上缓解模式崩塌并降低网络训练难度. 后来 Arjovsky 等人[34]提出的 WGANs (Wasserstein generative adversarial networks) 彻底解决了 GAN 训练不稳定的问题, 并同时使模式崩塌问题得到基本解决. 该方法主要对 GAN 做出了以下 4 点改进: 1) 去掉鉴别网络最后一层的 Sigmoid; 2) 损失函数不取 log; 3) 对更新后的参数进行截断; 4) 不使用基于动量的优化算法. 而与 WGANs 在同一年提出的 least squares GANs (LSGANs)[35]仅通过将 GAN 的目标函数由交叉熵损失换成最小二乘损失, 便同时解决了上述两个问题. LSGANs 中, 鉴别网络和生成网络的目标函数分别定义如下:

$$\min_{D} V_{\text{LSGANs}}(D)=\frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}(x)}[(D(x)-label_{\text{real}})^2]+\frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z))-label_{\text{fake}})^2] \tag{3}$$

$$\min_{G} V_{\text{LSGANs}}(G)=\frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z))-label_{\text{real}})^2] \tag{4}$$

其中, $label_{\text{real}}$ 和 $label_{\text{fake}}$ 分别表示真实图像与生成图像的标签.

综上所述, 本文方法选择采用 LSGANs 来完成融合网络的对抗训练. 在本文的 MR-T1/MR-T2 图像融合任务中, 输入数据的特征分布与融合后图像的特征分布具有一定程度的相似性, 这能够有效避免模式崩塌. 同时, 本文采用两个对称的鉴别网络分别与融合网络建立对抗关系, 且鉴别网络被设计为简单的二分类网络. 本文首先利用源图像和融合网络产生的融合图像分别对两个鉴别器进行训练, 使其具有区分源图像和融合图像的能力. 随后根据鉴别网络对融合图像的分类结果指导特征融合网络进行参数更新, 激励融合网络产生更加逼真且足以欺骗两个鉴别器的融合图像. 两个鉴别网络以同等的权重对特征融合网络进行约束, 最终保证不同模态的特征在融合图像中的保留程度大致相同.

2 本文方法

本文提出了一种基于预训练模型特征提取的双对抗融合网络来实现 MR-T1/MR-T2 图像的融合. 该网络由一个包含预训练卷积神经网络 (CNN) 的特征提取模块, 一个基于 DenseNet 的特征融合模块和两个辅助融合网络进行参数优化的对称鉴别网络模块组成. 如图 2 所示, 该网络首先将 MR-T1/MR-T2 图像输入特征提取模块, 然后将提取的不同模态特征在通道维度进行拼接, 最后将拼接好的特征一同输入融合网络以产生融合结果. 在模型优化阶段, 本文方法首先对两个鉴别网络进行训练, 使其具备一定的分类能力, 然后利用两个鉴别网络分别对融合网络产生的结果进行分类, 并同时将分类结果反馈给融合网络. 融合网络根据反馈信息不断优化自身参数, 以产生更加逼真的融合图像去欺骗鉴别器. 网络整体训练依照上述两个过程循环迭代, 直到两个鉴别网络都不能区分相应的源图像与融合图像时, 模型收敛, 训练过程结束.

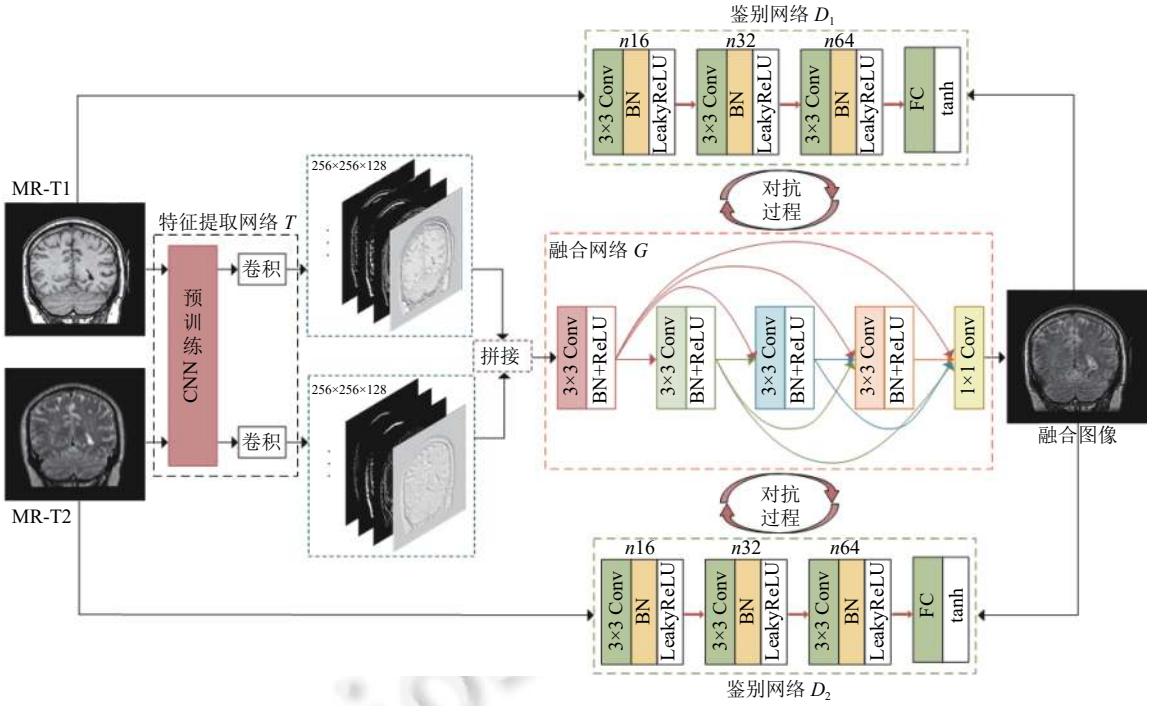


图 2 本文方法结构及流程图

2.1 问题描述

本文采用一个包含 K 对配准 MR-T1/MR-T2 图像的集合 $O=\{o_i\}_{i=1}^{K}$ 来实现网络的训练, 其中 $o_i=\{a_i,b_i\}$, $a_i\in\mathbb{R}^{W\times H\times C}$ 和 $b_i\in\mathbb{R}^{W\times H\times C}$ 分别代表源图像 MR-T1 与 MR-T2. W 和 H 表示图像的宽度和长度. C 表示源图像颜色通道的数量. 为进一步简化描述, 本文将 O 中不同模态的图像集合分别定义为 $a=\{a_1,a_2,\ldots,a_K\}\in\mathbb{R}^{K\times W\times H\times C}$, $b=\{b_1,b_2,\ldots,b_K\}\in\mathbb{R}^{K\times W\times H\times C}$, 并将所有融合图像定义为 $f=\{f_1,f_2,\ldots,f_K\}\in\mathbb{R}^{K\times W\times H\times C}$, 其中 f_i 是由源图像 a_i 和 b_i 融合而成的图像. 本文方法的目标是对预训练的神经网络模型参数进行微调, 使其能够充分提取适用于图像融合任务的特征. 同时通过双鉴别网络与特征融合网络之间的对抗学习, 不断激励融合网络进行参数更新, 拟合出适用于 MR-T1/MR-T2 融合任务的最优融合参数. 最终利用本文模型能够将 MR-T1 和 MR-T2 图像融合成特征丰富的高质量图像.

2.2 预训练 CNN

一般来说, 神经网络本身可被认为是特征提取器, 其中间层的映射表示可用来重构图像的显著特征. 但基于深度学习的特征提取方法一般需要大量专门的数据进行训练, 并需要占用大量的内存资源, 故本文采用预先在 ImageNet 上进行分类训练的 VGG-16 模型[36]分别对源图像 MR-T1 和 MR-T2 提取特征. 该 VGG 模型的卷积部分包含 5 个卷积块和 5 个下采样层, 其中前两个卷积块各包含两个卷积层, 后 3 个卷积块各包含 3 个卷积层. 由于 VGG 具有多个池化层, 在计算过程中会不断降低特征图分辨率, 为确保所得特征图的尺寸与源图像保持一致, 以便进行后续的融合操作, 本文使用该 VGG 模型的第一个卷积块对源图像进行特征提取. 如图 3 所示, 第 1 个卷积块中的两个卷积层均包含 64 个大小为 3×3 的卷积核, 足以提取广泛的低层特征. 除此之外, 为了进一步扩大特征感受野, 获取图像的更深层特征, 本文在该预训练模型之后增加一层卷积对提取的特征进行调整. 由于预训练的 VGG-16 是用于图像分类任务的, 所以在训练过程中, 本文通过随机梯度下降算法对现有的预训练参数进行微调, 保证预训练模型的参数适用于 MR-T1/MR-T2 图像融合任务. 此外, 我们定义 $f_k^{c,l}$ 为第 k 张图像在第 l 个卷积层计算得到的第 c 个特征图, 则该特征图的计算公式如下:

$$f_k^{c,l} = \max(0, F_l^c(I_k)) \tag{5}$$

其中, I_k 表示第 k 张图像. $F_l^c(\cdot)$ 代表第 l 个卷积层中第 c 个卷积核计算出的特征图. max(0,·)指的是 ReLU 激活函数.

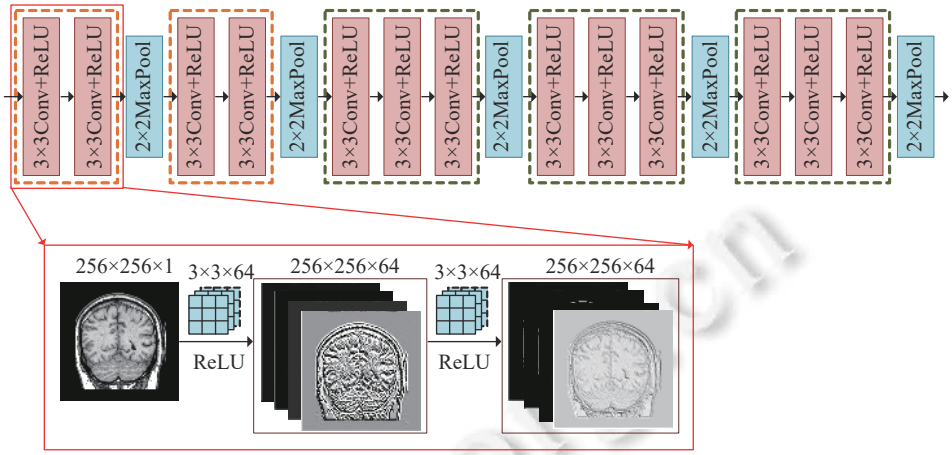


图 3 预训练模型网络结构

2.3 特征融合网络

目前基于深度学习的融合方法大多是基于 CNN 的, 且该类方法通常只将最后一层神经网络的输出作为融合结果, 这会导致大量中间层有用信息的丢失. 为此, 本文将特征融合网络 G 定义为具有 5 个卷积层的 DenseNet. 如图 2 所示, 该网络以前馈的方式在每一层和所有层之间建立短的直接连接, 进而弥补融合过程中的特征损失, 同时减轻梯度消失以及实现特征复用. 其中, 前 4 个卷积层用于特征融合, 最后一个卷积层实现特征降维并输出融合图像. 训练过程中, 前 4 层的卷积核大小均为 3×3, 移动步长为 1. 为保证输出特征图的大小不发生改变, 我们将 padding 设置为 1. 如图 4 所示, 在融合过程中, 每个卷积层首先对之前各层输出的特征图 ($\mathcal{F}_{k-2}$, $\mathcal{F}_{k-1}$, $\mathcal{F}_k$) 在通道维度进行拼接, 然后利用卷积核对拼接后的特征张量进行局部加权求和. 为保证特征的稀疏性, 避免融合模型过拟合, 本文最后通过 ReLU 激活函数将计算后的特征进行非线性映射即得到该层融合特征. 总体来说, 定义多个卷积层能引入更多参数, 并拟合出更复杂精确的融合函数. 此外, 本文通过在前 4 个卷积层后添加 batch-normalization[37] 来解决过拟合问题, 以增强训练过程的稳定性.

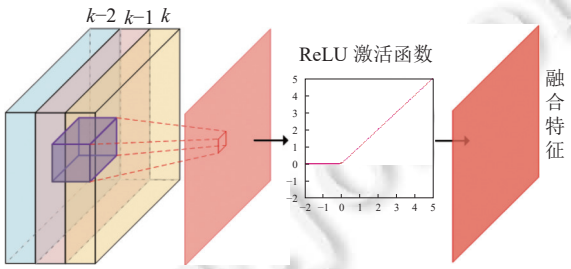
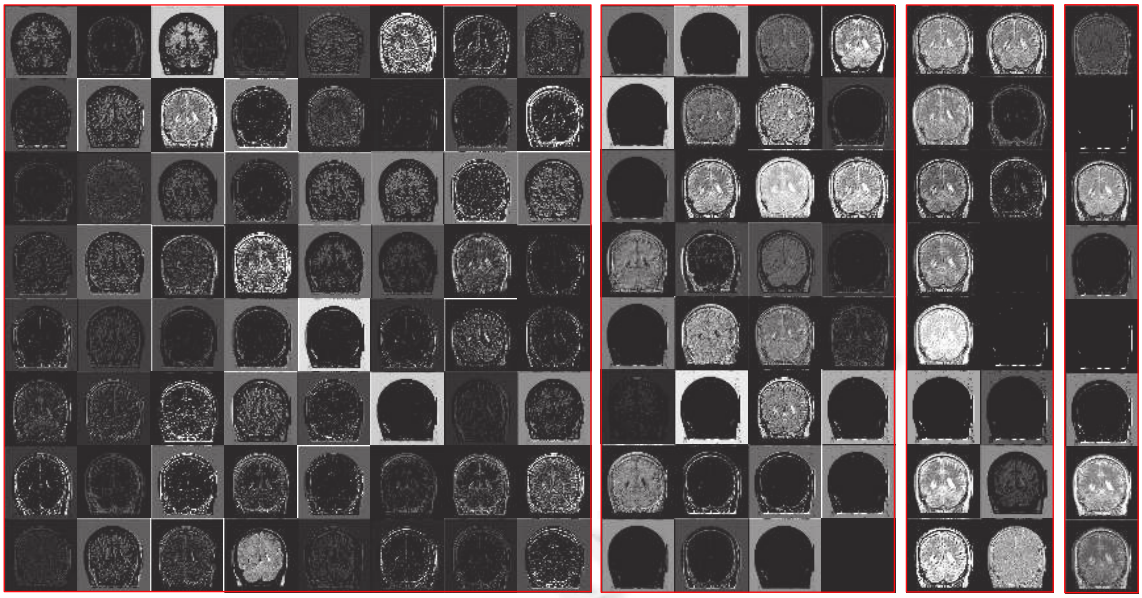


图 4 特征融合过程

此外, 为了能够更加清晰地对融合过程进行分析, 本文将特征融合网络中前 4 个卷积层的全部特征图在图 5 中进行了可视化. 其中, 图 5(a) 包含了第 1 层卷积产生的 64 个特征图, 图 5(b) 包含了第 2 层卷积产生的 32 个特征图, 图 5(c) 包含了第 3 层卷积产生的 16 个特征图, 图 5(d) 包含了第 4 层卷积产生的 8 个特征图. 如图 5 所示, 在融合过程中, 特征融合网络主要针对 MR-T1 和 MR-T2 的轮廓特征、像素分布特征以及各个方向的梯度特征等进行了重点保留与融合, 并且随着卷积次数的增加, 特征图中所包含的信息也越来越丰富.



(a) (b) (c) (d)

图 5 特征融合网络中间层特征图

2.4 鉴别网络

本文将两个鉴别网络均定义为具有 3 个卷积层和 1 个全连接层的分类网络, 且每层卷积核的大小均为 3×3. 为了避免在反向传播阶段中低层的神经网络发生梯度消失而导致模型无法进一步优化, 本文采用 batch-normalization 来减少不同 batch 间的数据抖动, 避免梯度消失并加速模型收敛. 同时, LeakyReLU 被作为鉴别网络的激活函数以实现特征数据的非线性映射, 避免激活函数梯度为零而导致模型无法收敛. 由于多模态医学图像融合的目的是对不同模态的信息进行整合, 因此应保证不同模态的特征在融合图像中能够得到大体同等程度地体现, 否则很容易出现融合图像过分表现某一模态的特征而淡化其他模态的特征. 鉴于此, 本文定义鉴别网络 D_1 和 D_2 具有相同的结构, 以保证二者对融合网络产生同等程度的对抗约束, 从而使得 MR-T1 和 MR-T2 的特征得到相同程度地保留. 总的来说, 正是由于双鉴别网络的对抗作用, 才使得融合网络能够在非监督的情况下自动实现对融合图像特征的充分表示与拟合.

训练过程中, 两个鉴别网络分别将融合网络产生的融合图像分类为假, 并根据分类结果利用反向传播激励融合网络进行参数更新, 以使其生成更加逼真的融合图像去欺骗鉴别网络. 当融合网络参数更新结束, 鉴别网络会再次根据新的融合图像和源图像进行训练以提高自身分类能力. 按照上述过程进行多次迭代, 直到鉴别网络无法将源图像和融合图像进行准确分类时, 训练过程结束, 此时融合网络学到最优的融合参数.

2.5 损失函数

损失函数由融合网络的损失函数和鉴别网络的损失函数两部分组成. 对融合网络而言, 本文不仅要对其进行对抗鉴别网络的训练, 还要对其生成的融合图像与源图像在内容相似性上进行约束. 因此, 融合网络的损失函数包括对抗损失和内容损失两部分. 同时, 由于预训练特征提取网络的参数微调过程是与特征融合网络参数的更新过程同步进行的, 因此特征融合网络的损失函数定义如下:

$$L_G(\theta_G,\theta_T)=L_{\text{adv}}(\theta_G,\theta_T)+L_{\text{con}}(\theta_G,\theta_T) \tag{6}$$

其中, θ_T 和 θ_G 分别表示特征提取网络与特征融合网络中的待更新参数. L_{adv} 代表融合网络与鉴别网络之间的对抗损失. L_{con} 表示用于增强细节特征和像素活动信息的内容损失. 对抗损失 L_{adv} 由融合网络与鉴别网络 D_1 以及融合网络与鉴别网络 D_2 两部分损失组成, 具体定义形式如下:

$$L_{\text{adv}}(\theta_G,\theta_T)=\frac{1}{K}\sum_{i=1}^{K}(D_1(f_i)-label_{r1})^2+\frac{1}{K}\sum_{i=1}^{K}(D_2(f_i)-label_{r2})^2 \tag{7}$$

内容损失L_{con}通过最小化均方误差重点对图像梯度和像素活动的相似性进行约束, 定义如下:

$$L_{\text{con}}(\theta_G,\theta_T)=\frac{1}{KHW}\sum_{i=1}^{K}(\alpha\|\nabla f_i-\nabla a_i\|_F^2+\beta\|f_i-b_i\|_F^2) \tag{8}$$

其中, ∇表示图像梯度. α和β是两个常数参数, 用于调节不同特征所占比重. 本文中, α和β的值将通过网格搜索进行分析与调整.

鉴别网络的对抗损失可以计算源图像与融合图像特征分布之间的相似性, 从而判断融合图像对源图像特征的保留程度. 本文方法的对抗损失由鉴别网络D_1的对抗损失和鉴别网络D_2的对抗损失两部分组成:

$$L_D(\theta_{D_1},\theta_{D_2})=L_{D_1}(\theta_{D_1})+L_{D_2}(\theta_{D_2}) \tag{9}$$

其中, $L_{D_1}(\theta_{D_1})$和$L_{D_2}(\theta_{D_2})$分别定义如下:

$$L_{D_1}(\theta_{D_1})=\frac{1}{K}\sum_{i=1}^{K}(D_1(a_i)-label_{r1})^2+\frac{1}{K}\sum_{i=1}^{K}(D_1(f_i)-label_f)^2 \tag{10}$$

$$L_{D_2}(\theta_{D_2})=\frac{1}{k}\sum_{i=1}^{k}(D_2(b_i)-label_{r2})^2+\frac{1}{k}\sum_{i=1}^{k}(D_2(f_i)-label_f)^2 \tag{11}$$

其中, $D(\cdot)\in[0,1]$表示鉴别网络的输出, 即输入图像的类别概率. $label_{r1}$和$label_{r2}$分别表示 MR-T1 和 MR-T2 的标签, $label_f$表示融合图像的标签.

2.6 参数更新

学习最优融合参数的过程是通过联合最小化融合网络的损失函数与鉴别网络的损失函数来进行的. 此外, 因融合网络与鉴别网络是两个相对独立的网络结构, 且二者的优化目标相反, 为保证训练的有序进行, 整个优化过程应分为两个交替的子过程:

$$(\widehat{\theta}_{D_1},\widehat{\theta}_{D_2})=\arg\min_{\theta_{D_1},\theta_{D_2}}(L_{D_1}(\theta_{D_1})+L_{D_2}(\theta_{D_2})) \tag{12}$$

$$(\widehat{\theta}_G,\widehat{\theta}_T)=\arg\min_{\theta_G,\theta_T}(L_{\text{adv}}(\theta_G,\theta_T)+L_{\text{con}}(\theta_G,\theta_T)) \tag{13}$$

上述两个子过程的优化可通过随机梯度下降算法来实现. 在一个迭代过程中, 本文方法首先根据当前融合图像与源图像对两个鉴别网络进行训练, 并更新其参数θ_{D_1}和θ_{D_2}, 直至两个鉴别器能够准确分辨出当前融合图像与源图像. 随后根据鉴别网络对当前融合图像的分类结果更新融合网络的参数θ_G以及特征提取网络的参数θ_T, 激励融合网络生成更加逼真且足以欺骗鉴别网络的融合图像. 网络具体优化过程如算法 1 所示.

算法 1. 本文方法参数优化伪代码.

初始化:

每个 minibatch 中某模态图像的数量: K;

当前 batch 中的 MR-T1 图像: $\boldsymbol{a}=\{a_1, a_2,\ldots, a_K\}$;

当前 batch 中的 MR-T2 图像: $\boldsymbol{b}=\{b_1, b_2,\ldots, b_K\}$;

超参数: epoch, α, β;

更新直至收敛:

1. **for** epoch **do**:

2. 通过随机梯度下降更新鉴别网络参数:

3. $\theta_{D_1}\leftarrow\theta_{D_1}-\nabla_{\theta_{D_1}}\frac{1}{K}\left(\sum_{i=1}^{K}((D_1(a_i)-label_{r1})^2+(D_1(f_i)-label_f)^2)\right)$;

4. $\theta_{D_2} \leftarrow \theta_{D_2} - \nabla_{\theta_{D_2}} \frac{1}{K}\left(\sum_{i=1}^{K}((D_2(b_i)-label_{r2})^2+(D_2(f_i)-label_f)^2)\right)$;
5. 固定鉴别网络的参数, 更新特征提取网络与特征融合网络的参数:
6. $L_{\text{adv}}(\theta_G,\theta_T) \leftarrow \frac{1}{K}\sum_{i=1}^{K}((D_1(f_i)-label_{r1})^2+(D_2(f_i)-label_{r2})^2)$;
7. $L_{\text{con}}(\theta_G,\theta_T) \leftarrow \frac{1}{KHW}\sum_{i=1}^{K}(\alpha\|\nabla f_i-\nabla b_i\|_F^2+\beta\|f_i-b_i\|_F^2)$;
8. $(\theta_G,\theta_T) \leftarrow (\theta_G,\theta_T)-\nabla_{(\theta_G,\theta_T)}(L_{\text{adv}}(\theta_G,\theta_T)+L_{\text{con}}(\theta_G,\theta_T))$;
9. **end for**;

return 最优生成器模型;

3 实验分析

实验基于来自 Harvard medical school website (http://www.med.harvard.edu/AANLIB/home.html) 的公开数据集对本文方法进行验证. 该数据集共包含 377 对配准的 MR-T1/MR-T2 图像, 每种模态的图像集合中均包括人类脑部横轴位、矢状位、冠状位 3 种不同角度的图像. 其中每张图像的大小均为 256×256×1 pt. 为了确保模型能够得到充分地训练, 本文随机选取了 20 个图像对用于测试拟合模型, 剩余的 357 个图像对全部用于模型训练. 训练过程中, 学习率被设置为经验值 3E–3, 衰减系数设置为 0.75. 受限于硬件设备的计算能力, 本文将 batch-size 设置为 2. 除此之外, 本文方法是基于 PyTorch 深度学习框架实现的. 本文方法产生的融合模型生成一张融合图像的平均时间为 1.23 s, 基本能够满足医学图像处理领域对实时性的要求.

本节首先介绍本文所用到的融合质量评价指标, 然后对超参数 α 、β 以及 epoch 的取值进行实验分析. 同时, 我们将本文网络中的预训练模型替换为具有相同结构的普通 CNN 网络, 在其他参数设置均不变的情况下进行消融实验, 并根据实验结果分析预训练技术在多模态医学图像融合任务中的优势. 除此之外, 我们通过将本文方法与 6 种现有的图像融合方法进行对比实验, 来证明本文方法的有效性. 上述实验均是在 40 GHz Intel Core i3-3240 CPU, GeForce GTX 1080 以及 32 GB 内存的计算平台上进行的.

3.1 评价指标

由于多模态医学图像融合是一类无监督的学习任务, 无法通过计算融合结果与 ground-truth 之间的损失来评判融合结果的优劣, 故实验选用了峰值信噪比 (peak signal-to-noise ratio, *PSNR*)[38]、结构相似性 (structural similarity, *SSIM*)[39]、互信息 (mutual information, *MI*)[40]、信息熵 (entropy, *EN*)[41]、空间频率 (spatial frequency, *SF*)[42]、标准差 (standard deviation, *SD*)[43]、视觉信息保真度 (visual information fidelity, *VIF*)[44]和 *Qabf*[45]这 8 种该领域常用的评价指标对融合结果的质量进行评价. 本文通过计算融合图像与源图像间的 *PSNR*、*SSIM* 和 *MI* 来评价二者之间的相似度. 指标的值越大表示源图像的特征在融合图像中的保留程度越高, 融合效果越好. 而 *EN*、*SF*、*SD* 和 *Qabf* 衡量的是融合图像的质量, 其值越高代表融合图像中所包含的信息越丰富, 融合图像质量越好. 其中, 指标 *VIF* 综合考虑了融合图像质量以及融合图像与源图像的相似度, 其值越高融合图像表现越好.

PSNR 通过计算两幅图像对应位置像素的均方误差 *MSE* 来评估两幅图像的相似性, 定义如下:

$$PSNR = 10\log\left(\frac{2^N-1}{MSE}\right)^2 \tag{14}$$

其中, N 表示每个像素所占的比特数. 本文中图像像素的取值范围是 [0, 255].

SSIM 通过计算源图像与融合图像的亮度损失, 对比度损失与结构损失来评价二者的相似度, 定义如下:

$$SSIM_{o,f} = \frac{2\mu_o\mu_f+C_1}{\mu_o^2+\mu_f^2+C_1}\cdot\frac{2\sigma_o\sigma_f+C_2}{\sigma_o^2+\sigma_f^2+C_2}\cdot\frac{\sigma_{of}+C_3}{\sigma_o\sigma_f+C_3} \tag{15}$$

其中, μ 与 σ 分别表示图像的均值和方差. C_1, C_2 和 C_3 是 3 个常数, 用于维持计算过程的稳定. 整体 *SSIM* 由融合图像与 MR-T1 的 *SSIM* 和融合图像与 MR-T2 的 *SSIM* 求和得出.

MI 通过计算源图像与融合图像的信息熵和联合信息熵来评估两者之间的相关性, 定义如下:

$$MI = E(f) + E(o) - E(f,o) \tag{16}$$

其中, $E(f)$ 和 $E(o)$ 分别表示融合图像与源图像的信息熵, $E(f,o)$ 代表融合图像与源图像的联合信息熵.

EN 用于评价融合图像中所包含的信息量, 定义如下:

$$EN = -\sum_{l=0}^{L-1} p_l \log_2 p_l \tag{17}$$

其中, L 代表灰度级, p_l 表示融合图像关于灰度级 l 的归一化直方图.

SF 通过分别计算图像的空间行频率和空间列频率来反映融合图像的梯度分布, 定义如下:

$$SF = \sqrt{\frac{\sum_{i=1}^{H}\sum_{j=1}^{W}[(I(i,j)-I(i,j-1))^2 + (I(i,j)-I(i-1,j))^2]}{HW}} \tag{18}$$

SD 基于统计原理反映图像中每个像素值与平均像素值之间的差异程度, 从而计算融合图像信息的丰富程度. 定义如下:

$$SD = \sqrt{\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(I(i,j)-\mu)^2} \tag{19}$$

Qabf 能够通过多个滑动窗口遍历整张图像, 计算出整张图像的边缘信息保留程度. *Qabf* 定义如下:

$$\begin{cases} Q_0(o,f) = \dfrac{\sigma_{of}}{\sigma_o\sigma_f}\cdot\dfrac{2\overline{of}}{\overline{o}^2+\overline{f}^2}\cdot\dfrac{2\sigma_o\sigma_f}{\sigma_o^2\sigma_f^2} \\ Qabf = \dfrac{1}{|W|}\displaystyle\sum_{w\in W} c(w)(\lambda(w)Q_0(a,f|w) + (1-\lambda(w))Q_0(b,f|w)) \end{cases} \tag{20}$$

其中, W 指滑动窗口集合. $Q_0(o,f)\in[-1,1]$ 计算的是融合图像与源图像的相似性. $Q_0(a,f|w)$ 与 $Q_0(b,f|w)$ 分别表示滑动窗口内 MR-T1、MR-T2 与融合图像的相似性. $\lambda(w)\in[0,1]$ 用于调整 $Q_0(a,f|w)$ 和 $Q_0(b,f|w)$ 所占比重. $c(w)$ 是根据图像的局部显著性计算出的权重.

VIF 计算的是融合图像的视觉信息保真度. 该指标的计算过程首先需要将目标图像分割成多个子块, 然后计算每个子块的 *VIF*, 评估其视觉信息是否失真. 最后, 将所有子块的 *VIF* 加和以获得整体图像的视觉信息保真度.

3.2 参数分析

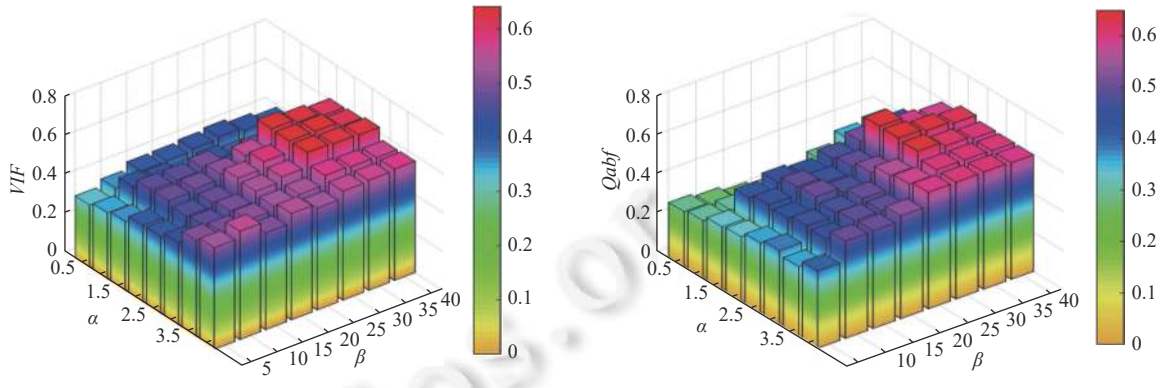
本文方法分别通过公式 (8) 中的参数 α 和 β 控制 MR-T1 中纹理与 MR-T2 中像素活动的保留程度, α 与 β 的不同取值会对融合结果产生一定程度的影响. 为此, 本文针对 α 与 β 的不同取值进行实验对比与分析, 并根据融合图像的视觉效果和量化指标最终确定 α 与 β 的值. 在参数分析过程中, 因指标 *VIF* 综合考虑了融合图像与源图像的相似度以及融合图像的质量, 所以被选作量化指标之一. 除此之外, 为重点分析融合图像边缘信息的保留程度, 本文还对 α 与 β 不同取值下 *Qabf* 的变化进行了展示. 在此过程中, 我们分别将 α 与 β 的取值范围设置为 {0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4} 和 {5, 10, 15, 20, 25, 30, 35, 40}. 如图 6 所示, 当 $\alpha\in[2,3]$, $\beta\in[30,40]$ 时, 指标 *VIF* 取得最高值, 这表明当 α 和 β 的值分别取自该范围时, 融合效果最好. 并且, 当 $\alpha\in[2,3]$, β=30 时, 融合图像的边缘信息保留度最高.

此外, 本文将融合网络与鉴别网络的损失函数在 1–120 个 epoch 上的收敛情况进行了可视化分析. 如图 7(a) 所示, 整个训练过程中, 在预训练模型的配合下, 融合网络与鉴别网络的损失几乎是单调递减并平稳收敛的, 且二者的损失均在 100 个 epoch 后收敛, 故本文方法的 epoch 被设置为 100. 上述所有参数设置均被用于接下来的对比实验.

3.3 消融实验

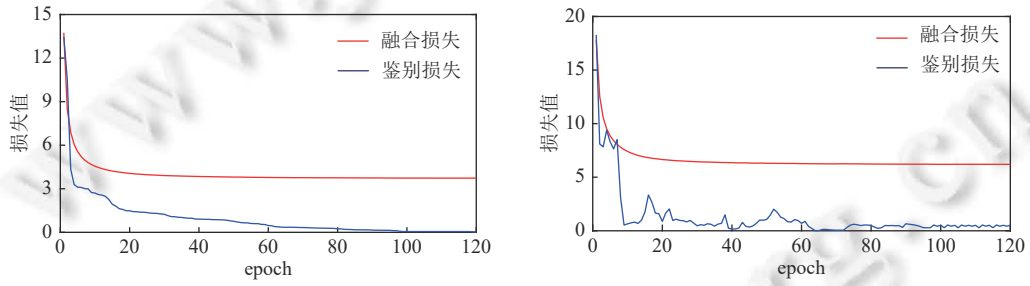
为了验证预训练模型在多模态医学图像融合任务中起到了关键作用, 本文对有预训练模型的网络和无预训练

模型的网络进行了消融实验与结果分析. 同时, 为排除网络结构或参数量等因素对实验结果的影响, 本文采用控制变量的方法, 令无预训练模型的网络在网络结构与参数设置上均与有预训练模型的网络相同 (如图 2). 即有预训练模型的网络和无预训练模型的网络不同之处在于: 前者特征提取模块使用预训练好的参数进行初始化; 后者特征提取模块参数的初始化方式为随机初始化. 首先本文对有预训练模型和无预训练模型的网络收敛情况分别进行了可视化. 根据图 7 可知, 在预训练模型的配合下, 网络的整体训练过程更加稳定, 且收敛速度加快. 此外, 相较于无预训练模型的网络, 本文网络在训练过程中的损失整体较低. 由此可见, 在网络中加入预训练模型, 能够辅助网络参数快速接近全局最优解, 并且能够有效提高网络的鲁棒性.



(a) *VIF* (b) *Qabf*

图 6 不同α、β取值下的指标 *VIF* 和 *Qabf* 变化情况



(a) 有预训练模型 (b) 无预训练模型

图 7 损失函数收敛变化情况对比

另外, 本文将有预训练模型网络和无预训练模型网络的融合结果进行了展示和对比. 由后文图 8 可以看出, 有预训练模型的网络产生的融合图像对比度更高, 图像的纹理细节更加清晰, 且高信号特征的分布更加明显. 除了视觉效果的展示, 本文还通过计算量化指标来评价两种融合结果的优劣. 如后文表 1 所示, 在 8 种常用的客观评价指标中, 有预训练模型网络产生的融合结果的表现普遍优于无预训练模型产生的融合结果, 与视觉表现基本一致.

3.4 对比实验

为进一步证明本文方法的有效性, 我们将本文方法与 ZLFMIF[22]、PMGI[21]、FusionGAN[24]、IFCNN[23]、PAPCNN[20]、NSCT[15]这 6 种典型方法进行了对比实验. 其中, NSCT 是比较经典的传统融合方法, 其余方法均为基于深度学习的图像融合方法. 部分 MR-T1/MR-T2 源图像及其相应的融合对比结果如图 8 所示, MR-T1 图像主要通过清晰的纹理反映脑部结构和区域信息, 而 MR-T2 图像主要通过不同的像素活动反映脑部组织或结构的状态. 因此, 本文方法除了利用对抗策略对源图像整体特征进行保留之外, 还通过定义内容损失函数对 MR-T1 的纹

理特征和 MR-T2 的像素强度分布进行了重点保留. 根据公式 (8), 该损失函数可通过最小化融合图像与源图像相关特征的误差来实现. 此外, 在计算纹理特征的均方误差之前, 本文先利用 Sobel 算子[46]分别获取 MR-T1 与融合图像的梯度. 根据图 9 可以看出, 在大部分的脑部区域中, MR-T1 与 MR-T2 的梯度变化是相反的. 为避免 MR-T2 的梯度对 MR-T1 纹理的保留产生影响, 本文利用 9×9 的滤波器对 MR-T2 和融合图像进行平滑处理, 然后针对平滑后的图像计算像素活动的均方误差.

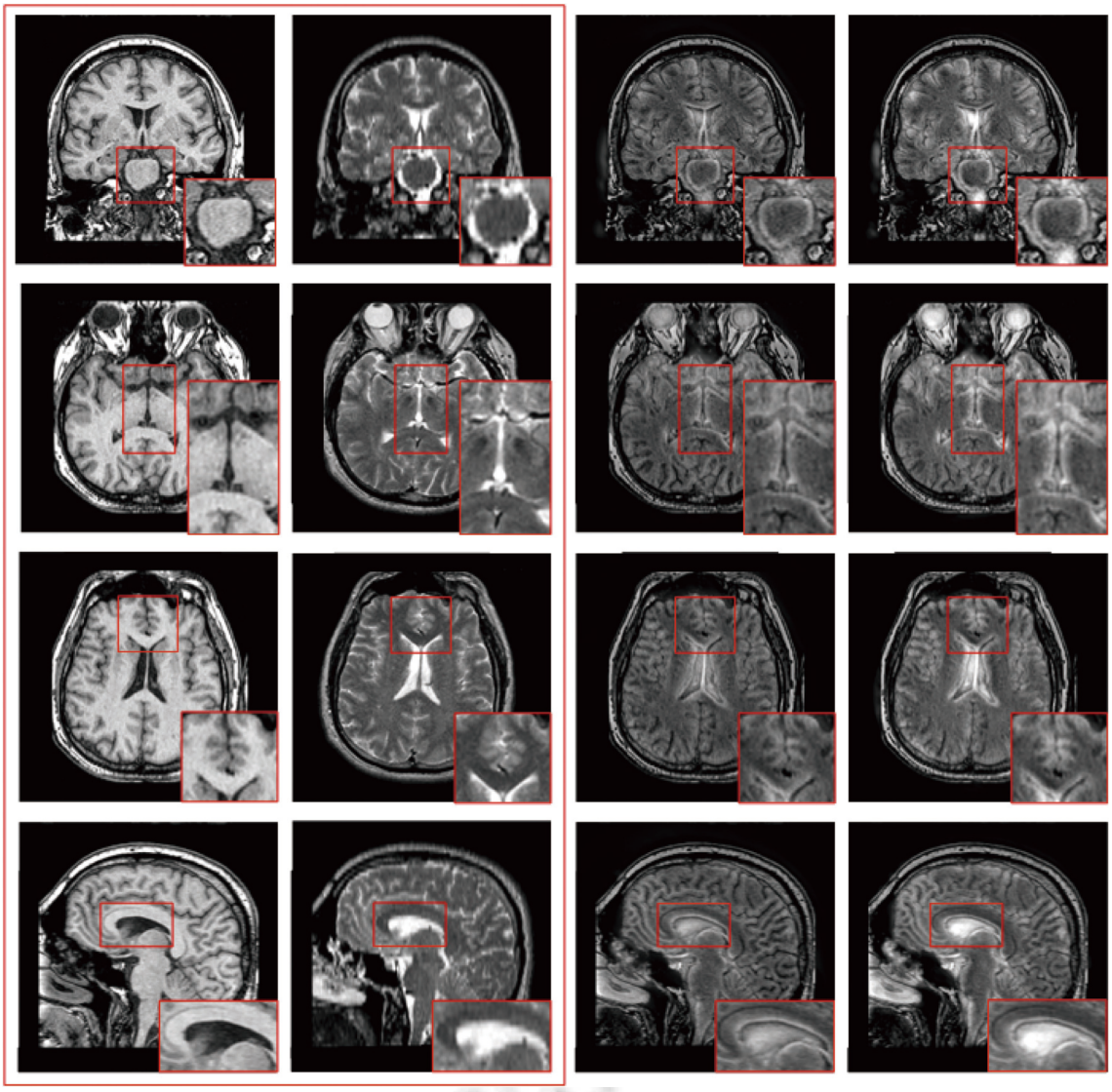


图 8 消融实验融合结果对比

表 1 消融实验融合结果平均量化指标对比

有/无预训练	*PSNR* (dB)	*SSIM*	*MI*	*EN* (bit)	*SF*	*SD* (pt)	*VIF*	*Qabf*
无预训练	17.3233	0.6874	1.1049	**5.6039**	0.1418	72.4254	0.6248	0.6192
有预训练	**18.0213**	**0.7176**	**1.2589**	5.5835	**0.1458**	**79.0338**	**0.6754**	**0.6482**

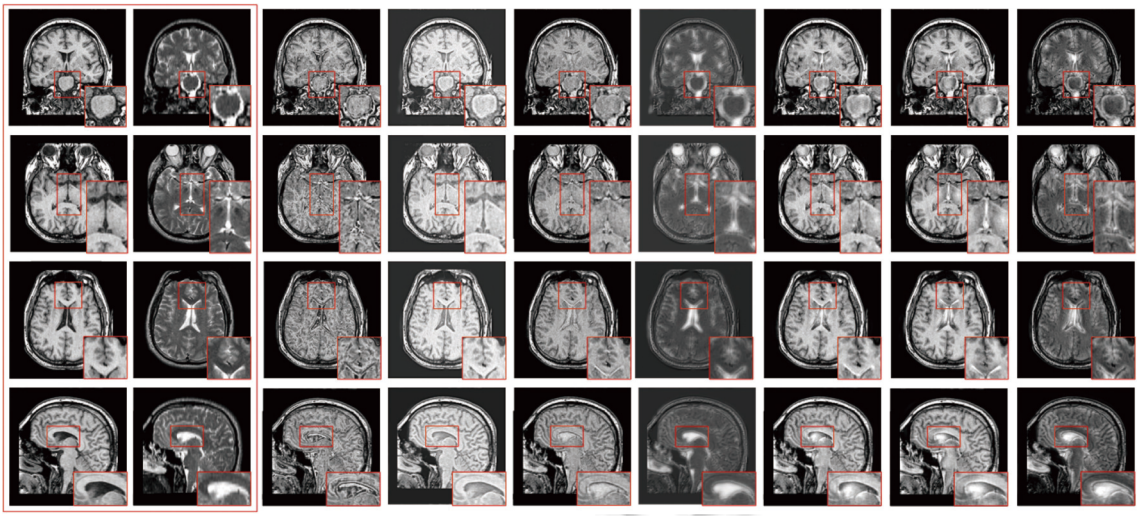


图 9 MR-T1/MR-T2 源图像以及各对比方法融合结果

如图 9 所示, PMGI、PAPCNN 和 NSCT 这 3 种方法的融合结果在视觉上都着重保留了 MR-T1 的纹理和亮度, 而与 MR-T2 相关的强信号特征在某些脑部结构上丢失严重. 例如在图 9 第 1 行放大的脑部结构中, MR-T2 中代表自由水的高亮信号在上述 3 种方法的融合图像中完全没有体现. 在 ZLFMIF 的融合结果中, 我们可以明显看出该方法试图将 MR-T1 的纹理和亮度特征以及 MR-T2 的强信号特征进行融合. 同时为了减轻 MR-T2 的梯度对 MR-T1 纹理的保留产生影响, 该方法对融合图像的纹理细节进行了增强, 但这也在一定程度上引入了噪声. IFCNN 方法较为完整地保留了 MR-T2 图像的强信号特征. 但由于该方法同时保留了 MR-T1 的亮度特征, 从而导致脑内液体与脂肪组织表现出相同的信号强度, 使得具有不同状态的组织之间不具有区分性. FusionGAN 方法中使用了一个鉴别器仅针对 MR-T2 特征的保留与生成器进行对抗训练. 相应地, 根据其融合结果可以看出, 该方法对 MR-T2 的显著特征进行了较为充分地保留, 但严重丢失了 MR-T1 中的纹理细节特征. 相比之下, 本文方法重点保留了 MR-T1 的纹理特征和 MR-T2 的像素强度分布特征, 而没有将图像的亮度保留作为重点. 实验结果表明, 相比于其他方法, 本文方法能够对 MR-T1 与 MR-T2 的显著特征进行最大程度的保留.

除此之外, 本文还对不同方法产生的融合结果进行了多种评价指标的量化对比. 为了便于展示, 本文对所得量化结果取平均值. 如表 2 所示, 本文方法在指标 *SSIM*、*EN*、*SF* 和 *VIF* 上均取得了最优值. 对于指标 *MI* 和 *Qabf* 来说, 本文融合图像均取得了次优结果且仅与最优值分别相差 0.043 2 和 0.001 1. 除此之外, 由于本文方法没有考虑融合图像的亮度, 所以会导致指标 *PSNR*、*SD* 的值有所下降. 为了进一步证明本文方法的有效性, 我们计算了融合图像与不同模态源图像的相似度指标. 如表 3 所示, 本文方法所得融合图像与 MR-T1 和 MR-T2 的相似度差距最小. 这表明在本文方法中, MR-T1 与 MR-T2 的特征得到了较为均衡地保留.

表 2 融合图像平均量化指标

方法	*PSNR* (dB)	*SSIM*	*MI*	*EN* (bit)	*SF*	*SD* (pt)	*VIF*	*Qabf*
ZLFMIF	16.516 5	0.659 0	**1.302 1**	5.127 0	0.144 1	78.711 8	0.591 4	0.628 6
PMGI	12.790 0	0.234 6	0.830 7	5.148 1	0.077 7	76.337 7	0.557 4	**0.649 3**
FusionGAN	12.637 6	0.292 0	1.093 9	5.068 3	0.124 0	59.480 7	0.163 9	0.158 2
IFCNN	17.617 5	0.712 5	1.083 9	5.193 5	0.124 7	77.254 4	0.651 2	0.589 6
PAPCNN	18.657 3	0.702 3	1.229 9	5.516 5	0.122 0	**84.232 5**	0.669 7	0.580 8
NSCT	**18.760 8**	0.694 1	1.238 4	5.456 2	0.126 5	83.495 8	0.665 9	0.642 6
本文方法	18.021 3	**0.717 6**	1.258 9	**5.583 5**	**0.145 8**	79.033 8	**0.674 5**	0.648 2

注: 加粗数字表示每列的最优(大)值

表 3 融合图像与源图像的相似度平均量化指标

指标	图像	ZLFMIF	PMGI	FusionGAN	IFCNN	PAPCNN	NSCT	本文方法
PSNR (dB)	MR-T1	20.7332	11.2534	15.9610	22.0162	25.6475	25.9262	16.6698
	MR-T2	12.2997	14.3266	9.3142	13.2187	11.6671	11.5953	19.3729
SSIM	MR-T1	0.8600	0.2514	0.5469	0.9114	0.9103	0.9376	0.8059
	MR-T2	0.4580	0.2178	0.0371	0.5135	0.4942	0.4505	0.6294
MI	MR-T1	2.0028	0.8179	1.4307	1.4107	1.7129	1.7128	1.3753
	MR-T2	0.6013	0.8435	0.7571	0.7571	0.7468	0.7640	1.1444

为进一步证明对称双鉴别网络对于 MR-T1 和 MR-T2 的特征都进行了融合保留, 本文分别计算了用于测试拟合模型的 20 个 MR-T1/MR-T2 图像对集合, 以及二者与其融合图像的 Fréchet Inception 距离 (Fréchet Inception distance, FID)[47]. FID 是专门用来衡量 GAN 模型生成图像的质量和多样性的指标. 该指标首先通过 Inception 网络来提取特征, 然后利用高斯模型对特征空间进行建模, 最后使用均值和协方差矩阵来求解真实样本与生成样本在特征空间上的距离. FID 越低表示两个分布之间的距离越小, 同时也表示生成图像的质量越高, 多样性越好. 其中 MR-T1 与 MR-T2、MR-T1 与融合图像、MR-T2 与融合图像的 FID 分别是 173.6961、95.5981、112.313 3. 由此可见, MR-T1 与 MR-T2 之间的 FID 相对较低, 表明 MR-T1 与 MR-T2 的特征分布具有一定程度的相似性. 同时, 相比于源图像本身的 FID, 融合图像与两种源图像之间的 FID 均有所降低, 且降低程度相差不大, 这说明融合图像对于 MR-T1 与 MR-T2 的分布进行了大致同等程度的拟合, 对不同模态的特征都进行了充分地保留.

4 总结与展望

为解决多模态医学图像融合数据规模小、无法充分训练模型且需人为设计约束易导致细节特征丢失等问题, 本文提出了一种预训练模型特征提取驱动下的双对抗多模态医学图像融合网络, 用于实现 MR-T1/MR-T2 图像的融合. 该网络首先利用预训练的 CNN 模型提取源图像特征, 随后对特征进行卷积以扩大特征感受野并获取深层特征. 然后, 不同模态的特征在通道维度进行拼接后被输入 DenseNet 融合网络进行加权融合. 本文模型在训练过程中, 主要通过两个鉴别网络与融合网络之间的双对抗学习过程实现非监督的多模态医学图像融合, 因而能够避免人工设计复杂的损失函数, 降低网络计算的复杂度. 此外, 本文还通过定义简单有效的内容损失函数增加对融合过程的约束, 以重点保留 MR-T1 的纹理和 MR-T2 的像素活动.

然而, 本文方法为了重点融合纹理与像素活动特征, 在一定程度上牺牲了融合图像的亮度, 这容易导致图像对比度有所下降. 因此, 在接下来的工作中, 本文将进一步针对 MR-T1/MR-T2 图像的特征分布展开研究. 同时, 我们将对本文方法进行改进, 通过引入注意力机制重点保留 MR-T1 和 MR-T2 的显著特征, 且不影响融合图像的亮度和对比度. 此外, 鉴于清华大学计算机系图形学实验室发布的开源 Jittor 深度学习框架 (https://cg.cs.tsinghua.edu.cn/jittor/download/) 在多种骨干网络模型中的动态图推理速度较 PyTorch 均有很大程度的提升, 因此我们将在未来工作中基于 Jittor 进行相关实验, 以提高网络的融合性能.

References:

[1] Liu H, Xu J, Wu Y, Guo Q, Ibragimov B, Xing L. Learning deconvolutional deep neural network for high resolution medical image reconstruction. Information Sciences. 2018, 468: 142–154. [doi: 10.1016/j.ins.2018.08.022]

[2] Liu H, Wang HO, Wu Y, Xing L. Superpixel region merging based on deep network for medical image segmentation. ACM Trans. on Intelligent Systems and Technology. 2020, 11(4): 1–22. [doi: 10.1145/3386090]

[3] Yu X, Liu H, Wu Y, Zhang CM. Intrinsic self-representation for multi-view subspace clustering. Scienta Sinica Informations, 2021, 51(10): 1625–1639 (in Chinese with English abstract). [doi: 10.1360/SSI-2020-0274]

[4] Hermessi H, Mourali O, Zagrouba E. Multimodal medical image fusion review: Theoretical background and recent advances. Signal Processing, 2021, 183: 108036. [doi: 10.1016/j.sigpro.2021.108036]

[5] Zhang H, Xu H, Tian X, Jiang JJ, Ma JY. Image fusion meets deep learning: A survey and perspective. Information Fusion, 2021, 76: 323–336. [doi: 10.1016/j.inffus.2021.06.008]

[6] Li Y, Zhao JL, Lv ZH, Li JH. Medical image fusion method by deep learning. Int'l Journal of Cognitive Computing in Engineering, 2021, 2: 21–29. [doi: 10.1016/j.ijcce.2020.12.004]

[7] Prabhakar KR, Srikar VS, Babu RV. DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 4714–4722. [doi: 10.1109/ICCV.2017.505]

[8] Li H, Wu XJ. DenseFuse: A fusion approach to infrared and visible images. IEEE Trans. on Image Processing, 2019, 28(5): 2614–2623. [doi: 10.1109/TIP.2018.2887342]

[9] Han X, Zhang ZY, Ding N, Gu YX, Liu X, Huo YQ, Qiu JZ, Yao Y, Zhang A, Zhang L, Han WT, Huang ML, Jin Q, Lan YY, Liu Y, Liu ZY, Lu ZW, Qiu XP, Song RH, Tang J, Wen JR, Yuan JH, Zhao WX, Zhu J. Pre-trained models: Past, present and future. AI Open, 2021, 2: 225–250. [doi: 10.1016/j.aiopen.2021.08.002]

[10] Iglovikov V, Shvets A. TernausNet: U-Net with VGG11 encoder pre-trained on ImageNet for image segmentation. arXiv:1801.05746, 2018.

[11] Chen YC, Li LJ, Yu LC, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu JJ. UNITER: Universal image-text representation learning. In: Proc. of the 2020 European Conf. on Computer Vision. Glasgow: Springer, 2020. 104–120. [doi: 10.1007/978-3-030-58577-8_7]

[12] Deng J, Dong W, Socher R, Li JL, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 20–25. [doi: 10.1109/CVPR.2009.5206848]

[13] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2261–2269. [doi: 10.1109/CVPR.2017.243]

[14] Sanjay AR, Soundrapandiyan R, Karuppiah M, Ganapathy R. CT and MRI image fusion based on discrete wavelet transform and type-2 fuzzy logic. Int'l Journal of Intelligent Engineering and Systems, 2017, 10(3): 355–362. [doi: 10.22266/ijies2017.0630.40]

[15] Bhatnagar G, Wu QMJ, Liu Z. Directive contrast based multimodal medical image fusion in NSCT domain. IEEE Trans. on Multimedia, 2013, 15(5): 1014–1024. [doi: 10.1109/TMM.2013.2244870]

[16] Zhang Q, Liu Y, Blum RS, Han JG, Tao DC. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. Information Fusion, 2018, 40: 57–75. [doi: 10.1016/j.inffus.2017.05.006]

[17] Bavirisetti DP, Xiao G, Liu G. Multi-sensor image fusion based on fourth order partial differential equations. In: Proc. of the 20th Int'l Conf. on Information Fusion. Xi'an: IEEE, 2017. 1–9. [doi: 10.23919/ICIF.2017.8009719]

[18] Han JG, Pauwels EJ, de Zeeuw P. Fast saliency-aware multi-modality image fusion. Neurocomputing, 2013, 111: 70–80. [doi: 10.1016/j.neucom.2012.12.015]

[19] Zheng YZ, Tan Z. Medical image fusion algorithm based on bidimensional empirical mode decomposition. Ruan Jian Xue Bao/Journal of Software, 2009, 20(5): 1096−1105 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3542.htm [doi: 10.3724/SP.J.1001.2009.03542]

[20] Yin M, Liu XN, Liu Y, Chen X. Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain. IEEE Trans. on Instrumentation and Measurement, 2019, 68(1): 49–64. [doi: 10.1109/TIM.2018.2838778]

[21] Zhang H, Xu H, Xiao Y, Guo XJ, Ma JY. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2019. 12797–12804.

[22] Lahoud F, Süsstrunk S. Zero-learning fast medical image fusion. In: Proc. of the 22nd Int'l Conf. on Information Fusion. Ottawa: IEEE, 2019. 1–8. [doi: 10.23919/FUSION43075.2019.9011178]

[23] Zhang Y, Liu Y, Sun P, Yan H, Zhao XL, Zhang L. IFCNN: A general image fusion framework based on convolutional neural network. Information Fusion, 2020, 54: 99–118. [doi: 10.1016/j.inffus.2019.07.011]

[24] Ma JY, Yu W, Liang PW, Li C, Jiang JJ. FusionGAN: A generative adversarial network for infrared and visible image fusion. Information Fusion, 2019, 48: 11–26. [doi: 10.1016/j.inffus.2018.09.004]

[25] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.

[26] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans. on Knowledge and Data Engineering, 2010, 22(10): 1345–1359. [doi: 10.1109/TKDE.2009.191]

[27] Raina R, Battle A, Lee H, Packer B, Ng AY. Self-taught learning: Transfer learning from unlabeled data. In: Proc. of the 24th Int'l Conf. on Machine Learning. Corvalis: Association for Computing Machinery, 2007. 759–766. [doi: 10.1145/1273496.1273592]

[28] Gao J, Fan W, Jiang J, Han JW. Knowledge transfer via multiple model local structure mapping. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Las Vegas: Association for Computing Machinery, 2008. 283–291. [doi: 10.1145/1401890.1401928]

[29] Du PF, Li XY, Gao YL. Survey on multimodal visual language representation learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 327–348 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[30] Hendrycks D, Lee K, Mazeika M. Using pre-training can improve model robustness and uncertainty. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 2712–2721.

[31] Chen FJ, Zhu F, Wu QX, Hao YM, Wang ED, Cui YG. A survey about image generation with generative adversarial nets. Chinese Journal of Computers, 2021, 44(2): 347–369 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2021.00347]

[32] Ratliff LJ, Burden SA, Sastry SS. Characterization and computation of local Nash equilibria in continuous games. In: Proc. of the 51st Annual Allerton Conf. on Communication, Control, and Computing. Monticello: IEEE, 2013. 917–924. [doi: 10.1109/Allerton.2013.6736623]

[33] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434, 2015.

[34] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv:1701.07875, 2017.

[35] Mao XD, Li Q, Xie HR, Lau RYK, Wang Z, Smolley SP. Least squares generative adversarial networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2794–2802. [doi: 10.1109/ICCV.2017.304]

[36] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[37] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: PMLR, 2015. 448–456.

[38] Huynh-Thu Q, Ghanbari M. Scope of validity of PSNR in image/video quality assessment. Electronics Letters, 2008, 44(13): 800–801. [doi: 10.1049/el:20080522]

[39] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans. on Image Processing, 2004, 13(4): 600–612. [doi: 10.1109/TIP.2003.819861]

[40] Qu GH, Zhang DL, Yan PF. Information measure for performance of image fusion. Electronics Letters, 2002, 38(7): 313–315. [doi: 10.1049/el:20020212]

[41] Roberts JW, van Aardt JA, Ahmed FB. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. Journal of Applied Remote Sensing, 2008, 2(1): 023522. [doi: 10.1117/1.2945910]

[42] Eskicioglu AM, Fisher PS. Image quality measures and their performance. IEEE Trans. on Communications, 1995, 43(12): 2959–2965. [doi: 10.1109/26.477498]

[43] Zhao ZB, Yuan JS, Gao Q, Kong YH. Wavelet image de-noising method based on noise standard deviation estimation. In: Proc. of the 2007 Int'l Conf. on Wavelet Analysis and Pattern Recognition. Beijing: IEEE, 2007. 1910–1914. [doi: 10.1109/ICWAPR.2007.4421768]

[44] Han Y, Cai YZ, Cao Y, Xu XM. A new image fusion performance metric based on visual information fidelity. Information Fusion, 2013, 14(2): 127–135. [doi: 10.1016/j.inffus.2011.08.002]

[45] Xydeas CS, Petrović V. Objective image fusion performance measure. Electronics Letters, 2000, 36(4): 308–309. [doi: 10.1049/el:20000267]

[46] Perra C, Massidda F, Giusto DD. Image blockiness evaluation based on Sobel operator. In: Proc. of the 2005 IEEE Int'l Conf. on Image Processing. Genova: IEEE, 2005. 386–389. [doi: 10.1109/ICIP.2005.1529769]

[47] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6629–6640.

附中文参考文献:

[3] 于晓, 刘慧, 吴彦, 张彩明. 基于本质自表示的多视角子空间聚类. 中国科学: 信息科学, 2021, 51(10): 1625–1639. [doi: 10.1360/SSI-2020-0274]

[19] 郑有志, 覃征. 基于二维经验模态分解的医学图像融合算法. 软件学报, 2009, 20(5): 1096–1105. http://www.jos.org.cn/1000-9825/3542.htm [doi: 10.3724/SP.J.1001.2009.03542]

[29] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述. 软件学报, 2021, 32(2): 327–348. http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[31] 陈佛计, 朱枫, 吴清潇, 郝颖明, 王恩德, 崔芸阁. 生成对抗网络及其在图像生成中的应用研究综述. 计算机学报, 2021, 44(2): 347–369. [doi: 10.11897/SP.J.1016.2021.00347]

刘慧(1978－), 女, 博士, 教授, 博士生导师, 主要研究领域为图像处理, 数据挖掘与可视化.

李珊珊(1997－), 女, 硕士生, 主要研究领域为机器学习, 多模态医学图像融合.

高珊珊(1980－), 女, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为智能图形图像处理, 数据挖掘与可视化.

邓凯(1981－), 男, 博士, 主任医师, 主要研究领域为医学图像诊断.

徐岗(1981－), 男, 博士, 教授, 博士生导师, 主要研究领域为智能图形图像处理, 几何计算与仿真.

张彩明(1955－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机图形学, 计算机视觉, 医学影像处理, 时序数据分析.