联合双支路生成对抗网络与 Transformer 的全色与多光谱遥感图像融合

姬云翔, 康家银, 马寒雁

江苏海洋大学 电子工程学院, 连云港 222005

摘 要：多光谱遥感图像具有能够反映丰富地物特征的光谱信息，但其空间分辨率较低，纹理信息相对不足。相反地，全色遥感图像的空间分辨率高，纹理信息丰富，但缺乏能够反映地物特征的丰富的光谱信息。通过图像融合技术可以将二者进行集成, 以达到各自的优势互补, 从而使得融合所得的图像能够更好地满足下游任务的需要。为此，本文提出了一种无监督的基于双支路生成对抗网络与 Transformer 的多光谱与全色遥感图像融合方法。具体地，首先采用引导滤波将源图像（源多光谱和全色遥感图像）分解为呈现图像主体信息的基础层分量与体现图像纹理、细节信息的细节层分量；然后，将分解得到的多光谱和全色遥感图像的基础层分量进行级联，将二者分解得到的细节层分量也进行级联；其次，将级联后的基础层分量和细节层分量分别输入至双支路生成器的基础层支路和细节层支路中；接着，针对基础层分量与细节层分量各自不同的特性，分别采用 Transformer 网络和卷积神经网络进行特征信息提取，以便从基础层分支和细节层分支中分别提取得到全局光谱信息和局部纹理信息；最后，通过生成器和双判别器（基础层判别器和细节层判别器）之间不断地对抗训练，得到同时具有丰富光谱信息与高空间分辨率的融合图像。通过在公开的数据集上与多个有代表性的方法进行定性与定量的对比实验表明，本文所提方法具有一定优越性，即在主观视觉效果和客观评价指标上均取得了较好的融合效果。

关键词：遥感图像融合；引导滤波；卷积神经网络；生成对抗网络；Transformer网络；基础层；细节层；全色；多光谱

中图分类号：TP391 **文献标志码**：A

收稿日期: 2009-01-06; **预印本:**2009-05-27

基金项目:国家自然科学基金面上项目（编号: 62271236）、江苏海洋大学自然科学基金项目（编号: Z2015009）、研究生科研与实践创新计划项目（编号: KYCX2022-41、KYCX2023-10）

第一作者简介: 姬云翔，1998 年生，男，硕士研究生，研究方向为图像处理与机器学习。E-mail: jyx990202@163.com

通信作者简介：康家银，1974 年生，男，教授，硕士生导师，主要研究方向为图像处理与机器学习。E-mail: kangjiayin2002@163.com

1 引言

随着遥感技术的发展，卫星等远距离传感器采集到的大量遥感图像在地质勘探、环境监测、城市规划、农业管理、灾害评估等领域得到广泛的应用 (Liu 等, 2018; Ma 等, 2019）。然而，由于采集设备的限制，同一传感器往往难以采集到同时具有丰富光谱信息和纹理信息的遥感图像，如在获得光谱分辨率较高的遥感图像时，传感器需要采集宽度较窄的波段信息，但这会导致其接收的辐射量减少，从而限制了采集到图像的空间分辨率等。因此，在实际应用中，单一传感器在获取遥感图像时需要在采集目标的光谱分辨率与空间分辨率之间做出折中，并针对不同需求采集具有不同信息的图像，如具有丰富光谱信息的多光谱图像 MS（Multispectral image）和具有清晰空间纹理信息的全色图像 PAN（Panchromatic image）等。为了弥补单一传感器获取图像信息的不足，通常采用图像融合技术从同一场景的不同图像中分别提取各自的优势信息，并尽可能地保留在生成的融合图像中 (Imani, 2019; Li 等, 2021）。如将多光谱图像与全色图像进行融合，可以得到同时具有丰富光谱信息与清晰空间纹理信息的融合图像。

现有的遥感图像融合研究方法，大致可以分为传统的融合方法和基于深度学习的融合方法（Zhang 等，2023）。传统的融合方法具有易于实现、可解释、处理速度快、便于针对不同场景任务进行优化等优点（Duran 等，2017）。具体地，传统的融合方法又可以进一步分为四类: 分量替换法 CS（Component Substitution）、多分辨率分析法 MRA（Multi-Resolution Analysis）、基于变分优化 VO（Variational Optimization）的方法以及基于模型的方法。CS 法首先对多光谱图像进行光谱变换，分离出保留丰富空间信息的成分分量；然后对全色图像与分离出的分量分别进行处理与替换；最后对其进行光谱逆变换得到融合结果（Garzelli 等，2007；Zhou 等，2014；Choi 等，2010）。分量替换法往往能够保持丰富的空间细节信息，但是会产生严重的光谱失真。MRA 法通常采取将全色图像进行分解并将分解

得到的高频空间细节信息整合到光谱波段中，常见的 MRA 法包括拉普拉斯金字塔 LP（Laplacian Pyramid）（Ranchin 等，2003）、离散小波变换 DWT（Discrete Wavelet Transform）（Pradhan 等，2006）、轮廓波变换 CT（Contourlet Transform）（Yang 等，2010）。VO 法往往将融合问题看作一个保留光谱信息和空间信息的代价函数的优化问题，而优化问题的解就是最终的融合图像（Ballester 等，2006）。基于模型的融合方法中，往往通过数据分布或数据表示模型来得到融合结果。Deng 等（2018）提出基于张量的稀疏模型和超拉普拉斯先验，在全色图像与多光谱图像融合问题上取得了较好的效果。传统方法往往需要人为设计复杂的融合规则以及先验，但是在实际中遥感图像融合无法转换成线性问题，导致融合结果通常具有严重的光谱失真问题。

在遥感图像融合中，最主要的问题是在融合的过程中减少空间纹理信息、光谱信息的丢失。由于深度学习强大的特征提取和数据处理能力，近年来基于深度学习的方法倍受研究者们的广泛关注，并已经在图像融合任务中取得了优异的性能。基于深度学习的图像融合方法有以下优势：1）深度学习模型可以从输入的数据中自动提取出最关键的数据特征，从而解决了人工设计特征难度较大的问题；2）深度学习模型可以很好地反映出输入数据与目标任务间复杂的映射关系；3）深度学习的一些潜在的图像表示方法更契合于图像融合任务；4）很多深度学习库和大规模图像数据集为基于深度学习的图像融合研究提供了帮助（Thomas 等，2008）。近期用于图像融合的深度学习方法主要包括卷积神经网络 CNN（Convolutional Neural Network），生成对抗网络 GAN（Generative Adversarial Network），深度残差网络 DRN（Deep Residual Network），Transformer 等。卷积神经网络主要通过利用卷积模板随网络层数的加深不断地学习图像不同层级的特征，然后将这些特征融合到一起，最终完成对图像的深层理解以达到融合图像的目的。该方法不需要手工设计和提取特征，能够自动学习图像数据的特征，因

此具有良好的适应性和泛化性能。卷积神经网络在遥感图像融合中的应用主要分为两类：一类是直接使用卷积神经网络对多源遥感图像进行融合；另一类是将卷积神经网络应用于特定的遥感图像融合算法中。在直接使用卷积神经网络进行遥感图像融合的方法中，研究者们通常采用多尺度融合的方法，即使用不同的卷积核对多源遥感图像进行卷积，并将不同尺度的卷积结果进行融合。Yang 等（2018）设计了一种双支路的卷积结构，针对多光谱图像与高光谱图像的特性分别进行特征提取并得到具有两种不同图像特征的较好融合结果。Shao 等（2018）则采用双支路的卷积结构针对多光谱图像与全色图像进行多尺度融合。另一种策略是将卷积神经网络应用于特定的遥感图像融合算法中，如 Ma 等（2021）提出了一个基于张量分解的低秩模型，通过结合图正则化来融合高光谱和多光谱图像，提高算法的性能和效果。GAN 方法通过采取对抗学习策略实现图像融合。GAN 的框架结构包含两个模型：一个是捕获数据分布的生成器；另一个是估计样本来自训练数据的概率的判别器。GAN 的本质是在生成器与判别器之间建立一个对抗博弈，并在不断地迭代中提升各自的性能。Ma 等（2019）首次将生成对抗网络用于图像融合，相较于传统融合方法，取得了较好的融合效果。在遥感图像融合领域中，Jin 等（2020）采用了条件生成对抗网络对全色图像与多光谱图像进行特征提取并进行融合。Ma 等（2020）设计了具有双判别器的生成对抗网络，从全色图像的空间纹理信息与多光谱图像的光谱信息两个层面对源图像进行特征提取并融合，最终生成出具有较好的空间纹理信息与光谱信息的融合图像。深度残差网络利用不同残差块组成的残差网络对输入的遥感图像进行特征学习，从而得到更精细的特征图，再将这些特征图融合到一起得到高质量的融合图像。该方法能够有效地处理遥感图像中的多尺度和多频带信息，提高融合图像的质量。Han 等（2019）采用深度残差网络对低分辨率高光谱图像与高分辨率多光谱图像进行特征提取并得到了具有较高分辨率的高光谱融

合图像。Qiu 等（2019）设计了一种新的双残差密集网络结构，在低分辨率高光谱图像与高分辨率多光谱图像的融合任务中实现了更好的效果。

Transformer 是一种广泛应用于自然语言处理的神经网络模型。近年来，Transformer 被广泛地应用于计算机视觉领域，并取得了较好的全局信息提取效果。基于 Transformer 的模型在遥感图像融合任务中的应用主要是通过自注意力机制来实现、并提高融合图像的质量（Dong 等，2021）。自注意力机制可以将不同尺度、不同位置的特征进行加权融合，从而最终提高融合图像的质量。Dosovitskiy 等（2021）首次将 Transformer 网络结构应用到图像处理领域，并提出了用于图像分类的 ViT(Vision Transformer)模型。Rao 等（2023）针对红外图像与可见光图像的融合任务，设计了将深度残差网络、Transformer 网络与生成对抗网络相结合的新型模型，从空间 Transformer 与通道 Transformer 两个角度对输入图像进行特征提取并融合，实现了较好的融合效果。Tang 等（2022）将 Transformer 网络与卷积神经网络结合，设计了一种“Y”字型结构的编码器-解码器融合框架，实现了红外图像与可见光图像较好的融合效果。针对目前基于 Transformer 网络的全色与多光谱遥感图像融合问题，考虑到具有清晰空间纹理信息的全色图像和具有丰富光谱信息的多光谱图像的不同特性，如何在网络中将 Transformer 较好的全局信息提取能力与 CNN 较好的局部信息提取能力进行有效结合以实现更好的融合效果仍是一个有待解决的问题。

基于以上分析，本文研究提出了一种无监督的基于双支路生成对抗网络和 Transformer 的遥感图像融合算法。本文的主要贡献体现在以下几个方面：1）本文提出了一种以多模态遥感图像经引导滤波分解后的不同层面分量为输入的双判别器生成对抗网络框架；2）本文设计了一种双支路生成器网络，即根据多模态遥感图像的基础层分量与细节层分量的不同特性，分别设计了基于卷积神经网络与基于 Transformer 网

络的两条生成器的网络支路；3）针对多模态遥感图像基础层分量与细节层分量的不同特点，分别设计了对应的损失函数；4）在公开的数据集上的实验结果表明，本文提出的算法不仅能充分地保留多光谱图像中的光谱信息，还能有效地集成全色图像中的空间纹理信息。

2 本文方法

本节首先展示本文所提的基于双支路生成对抗网络和 Transformer 的全色与多光谱遥感图像融合算法的总体框架；接着详细介绍网络训练过程中所使用的损失函数；最后详细介绍本文方法框架中卷积神经网络与 Transformer 网络部分的网络结构。

2.1 总体框架

现有的基于深度学习的全色与多光谱遥感图像融合方法中，较广泛使用的网络模型有基于CNN的网络模型和基于GAN的网络模型。基于 CNN 的网络模型要求有一个关键的先决条件，即需要事先获得真值(Ground Truth)以用于训练深度学习模型。实际中，全色与多光谱图像的融合结果并不存在真值，对融合图像的评判往往依靠人眼主观的视觉评价并辅以客观的评价指标。因此，基于 CNN 的网络模型在实际应用中会受到限制，从而影响其融合结果。相对地，基于 GAN 的网络模型训练时不依靠融合结果的真值，而是通过生成器与判别器之间的对抗博弈来达到训练的目的，因此更适合遥感图像的融合任务。

多光谱图像包含多个波段的信息，但由于噪声和其他因素的影响，这些信息可能被忽略或者难以分辨。引导滤波分解可以将不同波段的信息分解成具有丰富全局光谱信息的基础层（平滑分量）和具有清晰局部纹理信息的细节层（细节分量），从而使得多光谱图像中的不同信息处理更具有针对性。主成分分析可以将原始的多光谱图像转化为一组互不相关的主成分，其中每个主成分都包含了原始图像中的一部分信息。因此，在多光谱图像的分解过程中，采用其主成分分析图作为引导图像，可以去除多光谱图像中的冗余信息，提取出主要的特征信息，从而更好地指导滤波分解的过程。综上，本文

所提方法的总体框架如图 1 所示。其中，因全色图像尺寸与多光谱图像尺寸不一致，本文在数据预处理中对多光谱图像进行四倍上采样操作，以保持输入网络时源图像（全色图像、多光谱图像）的尺寸一致。

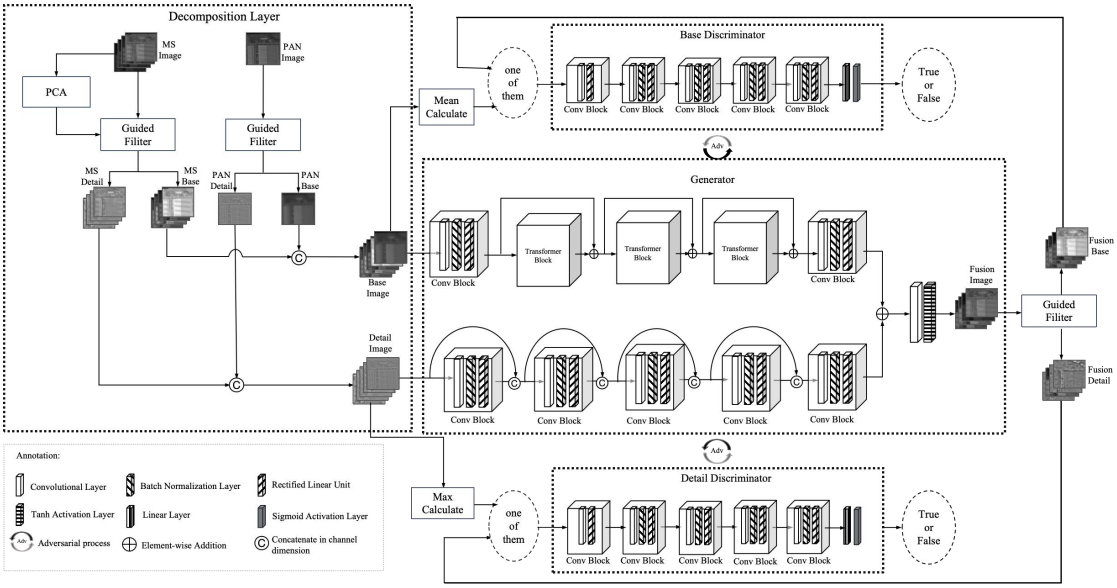


图 1 所提方法的总体框架图

Fig.1 Overall framework of the proposed method

在预处理中将多光谱图像进行四倍上采样处理，保持多光谱图像与全色图像尺寸一致，并同时输入网络。具体地，1)首先，使用引导滤波器将源图像进行分解，分别得到多光谱基础层、细节层与全色基础层、细节层；将多光谱基础层与全色基础层在通道维度堆叠，得到输入基础层图像；将多光谱细节层与全色细节层在通道维度堆叠，得到输入细节层图像。需要指出的是，在分解过程中，针对全色图像与多光谱图像各自不同的特性，分别对其采取不同的分解策略：全色图像的空间分辨率较高，具有清晰的纹理特征，故将其自身作为引导滤波器的引导图像；而多光谱图像的光谱分辨率高、光谱信息丰富但其纹理特征较为模糊，为防止出现其基础层过于模糊、轮廓难以分辨的情况，以及加强其分解效果，故采用对多光谱图像进行主成分分析并将其主成分图像作为引导图像的策略。2)其次，对于主要包含纹理信息的细节层图像，使用卷积神经网络实现其纹理信息的特征提取。3)此外，对于仍含有部分纹理细节信息以及较多光谱信息的

基础层图像，使用具有很强全局信息提取能力的 Transformer 网络进行特征提取，并将卷积神经网络提取的特征与 Transformer 网络提取的特征进行融合，然后基于融合的特征重构得到具有丰富纹理信息与光谱信息的融合图像，以此作为生成器的输出结果。4)接着，将生成器输出的融合图进行引导滤波分解，得到融合图像的基础层图像与细节层图像，并将其基础层图像与经过平均值计算的源图像基础层、其细节层图像与经过最大值计算的源图像细节层分别输入到基础层判别器、细节层判别器中，以便从基础层信息、细节层信息两个层面对输入的融合图像与源图像进行判别。5)最后，生成器与两判别器（基础层判别器、细节层判别器）不断地对抗与优化训练，直到基础层判别器与细节层判别器均无法辨别出生成器生成的融合图像时，该融合图像即为最终的融合结果。

简言之，本文所提的基于双支路生成对抗网络的遥感图像融合算法总体上主要从基础层与细节层两个层面入手，分别利用 Transformer 和 CNN 从中提取全局光谱信息与局部纹理信息，并将提取到的全局和局部信息整合到融合图像中，从而使得最终的融合图像既具有多光谱图像丰富的光谱信息，又包含全色图像清晰的纹理等细节信息。

2.2 损失函数

本文算法的损失函数由两部分组成：生成器损失和判别器损失。

2.2.1 生成器损失

生成器损失由对抗损失、内容损失与光谱损失三部分组成。生成器总损失定义如下：

$$L_{\mathrm{G}} = L_{\mathrm{adv}} + \lambda_1 L_{\mathrm{content}} + \lambda_2 L_{\mathrm{spectrum}} \tag{1}$$

其中 L_{G} 表示生成器 G 的总损失，λ_1、λ_2 为权重系数，L_{adv} 表示生成器 G 与基础层判别器 D_{base}、细节层判别器 D_{detail} 之间的对抗损失，具体定义如式(2)所示：

$$\begin{aligned} L_{\mathrm{adv}} &= \mathrm{E}\left[\log\left(1 - aD_{\mathrm{base}}\left(I_{\mathrm{F_{base}}}\right)\right)\right] \\ &+ \mathrm{E}\left[\log\left(1 - (1-a)D_{\mathrm{detail}}\left(I_{\mathrm{F_{detail}}}\right)\right)\right] \end{aligned} \tag{2}$$

其中 $I_{\mathrm{F_{base}}}$ 表示生成器生成的融合图像 I_{F} 的基础层图像，$I_{\mathrm{F_{detail}}}$ 表示 I_{F} 的细节层图像；a 为平衡 D_{base} 与 D_{detail} 的权重系数。

式(1)中生成器总损失的第二项 L_{content} 表示融合图像的内容损失，具体定义如下所示：

$$L_{\text{content}}=\alpha L_{\text{int}}+\beta L_{\text{grad}}+\gamma L_{\text{SSIM}} \tag{3}$$

其中 L_{int} 为强度损失，L_{grad} 为梯度损失，L_{SSIM} 为结构相似性损失；α 、β 、γ 为平衡三者的权重系数。L_{int} 的定义如下：

$$L_{\text{int}}=\frac{1}{HW}\left(\omega L_{\text{int-base}}+(1-\omega)L_{\text{int-detail}}\right) \tag{4}$$

上式中 H 、W 表示输入图像的高和宽；ω 为平衡两项的系数；$L_{\text{int-base}}$ 表示融合图像基础层 $I_{\text{F}_{\text{base}}}$ 与全色图像基础层 $I_{\text{Pan}_{\text{base}}}$ 、多光谱图像基础层 $I_{\text{MS}_{\text{base}}}$ 之间的基础层强度损失，具体定义如式(5)所示；$L_{\text{int-detail}}$ 表示融合图像细节层 $I_{\text{F}_{\text{detail}}}$ 与全色图像细节层 $I_{\text{Pan}_{\text{detail}}}$ 、多光谱图像细节层 $I_{\text{MS}_{\text{detail}}}$ 之间的细节层强度损失，具体定义如式(6)所示。

$$\begin{aligned}L_{\text{int-base}}=&b\left\|I_{\text{F}_{\text{base}}}-I_{\text{Pan}_{\text{base}}}\right\|_{\text{F}}\\&+(1-b)\left\|I_{\text{F}_{\text{base}}}-I_{\text{MS}_{\text{base}}}\right\|_{\text{F}}\end{aligned} \tag{5}$$

$$\begin{aligned}L_{\text{int-detail}}=&b\left\|I_{\text{F}_{\text{detail}}}-I_{\text{Pan}_{\text{detail}}}\right\|_{\text{F}}\\&+(1-b)\left\|I_{\text{F}_{\text{detail}}}-I_{\text{MS}_{\text{detail}}}\right\|_{\text{F}}\end{aligned} \tag{6}$$

在(5)、(6)两式中，b 均为平衡两项的权重系数；$\|\cdot\|_{\text{F}}$ 为 F 范数。

式(3)中内容损失 L_{content} 的第二项梯度损失 L_{grad} 的具体定义如下：

$$L_{\text{grad}}=\frac{1}{HW}\left[\omega L_{\text{grad-base}}+(1-\omega)L_{\text{grad-detail}}\right] \tag{7}$$

其中 $L_{\text{grad-base}}$ 表示融合图像基础层 $I_{\text{F}_{\text{base}}}$ 与全色图像基础层 $I_{\text{Pan}_{\text{base}}}$ 、多光谱图像基础层 $I_{\text{MS}_{\text{base}}}$ 之间的基础层梯度损失，具体定义如式(8)所示；$L_{\text{grad-detail}}$ 表示融合图像细节层 $I_{\text{F}_{\text{detail}}}$ 与全色图像细节层 $I_{\text{Pan}_{\text{detail}}}$ 、多光谱图像细节层 $I_{\text{MS}_{\text{detail}}}$ 之间的细节层梯度损失，具体定义如式(9)所示。

$$\begin{aligned}L_{\text{grad-base}}=&c\left\|\nabla I_{\text{F}_{\text{base}}}-\nabla I_{\text{Pan}_{\text{base}}}\right\|_{\text{F}}\\&+(1-c)\left\|\nabla I_{\text{F}_{\text{base}}}-\nabla I_{\text{MS}_{\text{base}}}\right\|_{\text{F}}\end{aligned} \tag{8}$$

$$\begin{aligned}L_{\text{grad-detail}}=&c\left\|\nabla I_{\text{F}_{\text{detail}}}-\nabla I_{\text{Pan}_{\text{detail}}}\right\|_{\text{F}}\\&+(1-c)\left\|\nabla I_{\text{F}_{\text{detail}}}-\nabla I_{\text{MS}_{\text{detail}}}\right\|_{\text{F}}\end{aligned} \tag{9}$$

在(8)、(9)两式中，c 均为平衡两项的权重系数。

式(3)中内容损失 L_{content} 的第三项结构相似性损失 L_{SSIM} 的具体定义如下：

$$L_{\text{SSIM}}=\omega L_{\text{SSIM-base}}+(1-\omega)L_{\text{SSIM-detail}} \tag{10}$$

其中 $L_{\text{SSIM-base}}$ 表示融合图像基础层 $I_{\text{F}_{\text{base}}}$ 与全色图像基础层 $I_{\text{Pan}_{\text{base}}}$ 、多光谱图像基础层 $I_{\text{MS}_{\text{base}}}$ 之间的基础层结构相似性损失，具体定义如式(11)所示；$L_{\text{SSIM-detail}}$ 表示融合图像细节层 $I_{\text{F}_{\text{detail}}}$ 与全色图像细节层 $I_{\text{Pan}_{\text{detail}}}$ 、多光谱

图像细节层 $I_{\mathrm{MS_{detail}}}$ 之间的细节层结构相似性损失，具体定义如式(12)所示。

$$\begin{aligned} L_{\text{SSIM-base}} &= \left(1 - L_{\text{SSIM}}\left(I_{\mathrm{F_{base}}}, I_{\mathrm{Pan_{base}}}\right)\right) \\ &+ \left(1 - L_{\text{SSIM}}\left(I_{\mathrm{F_{base}}}, I_{\mathrm{MS_{base}}}\right)\right) \end{aligned} \tag{11}$$

$$\begin{aligned} L_{\text{SSIM-detail}} &= \left(1 - L_{\text{SSIM}}\left(I_{\mathrm{F_{detail}}}, I_{\mathrm{Pan_{detail}}}\right)\right) \\ &+ \left(1 - L_{\text{SSIM}}\left(I_{\mathrm{F_{detail}}}, I_{\mathrm{MS_{detail}}}\right)\right) \end{aligned} \tag{12}$$

其中 $L_{\text{SSIM}}(\cdot)$ 表示两项的结构相似性。

式(1)中生成器总损失的第三项 L_{spectrum} 表示光谱损失，具体定义如下：

$$L_{\text{spectrum}} = L_{\text{spectrum-base}} + L_{\text{spectrum-detail}} \tag{13}$$

其中 $L_{\text{spectrum-base}}$ 表示多光谱图像基础层与融合图像基础层之间的基础层光谱损失，具体定义如式(14)所示；$L_{\text{spectrum-detail}}$ 表示多光谱图像细节层与融合图像细节层之间的细节层光谱损失，具体定义如式(15)所示：

$$L_{\text{spectrum-base}} = I - \frac{\left\langle I_{\mathrm{MS_{base}}} \cdot I_{\mathrm{F_{base}}} \right\rangle}{\left\| I_{\mathrm{MS_{base}}} \right\|_2 \left\| I_{\mathrm{F_{base}}} \right\|_2} \tag{14}$$

$$L_{\text{spectrum-detail}} = I - \frac{\left\langle I_{\mathrm{MS_{detail}}} \cdot I_{\mathrm{F_{detail}}} \right\rangle}{\left\| I_{\mathrm{MS_{detail}}} \right\|_2 \left\| I_{\mathrm{F_{detail}}} \right\|_2} \tag{15}$$

上式中，I 均为全 1 矩阵；$\langle \cdot \rangle$ 为两项内积；$\| \cdot \|_2$ 为 2 范数。

2.2.2 判别器损失

本文算法的判别器损失 L_{D} 由基础层判别器损失 $L_{\mathrm{D_{base}}}$ 和细节层判别器损失 $L_{\mathrm{D_{detail}}}$ 两部分组成。基础层判别器的损失定义如下：

$$\begin{aligned} L_{\mathrm{D_{base}}} &= \mathrm{E}\left[-\log\left(D_{\text{base}}(I_{\text{base-mean}})\right)\right] \\ &+ \mathrm{E}\left[-\log\left(1 - D_{\text{base}}(I_{\mathrm{F_{base}}})\right)\right] \end{aligned} \tag{16}$$

其中 D_{base} 表示基础层判别器，$D_{\text{base}}(\cdot)$ 表示基础层判别器对输入图像真假的判断值；$I_{\text{base-mean}}$ 表示多光谱图像基础层与全色图像基础层的进行平均值处理得到的图像。细节层判别器损失如下所示：

$$\begin{aligned} L_{\mathrm{D_{detail}}} &= \mathrm{E}\left[-\log\left(D_{\text{detail}}(I_{\text{detail-max}})\right)\right] \\ &+ \mathrm{E}\left[-\log\left(1 - D_{\text{detail}}(I_{\mathrm{F_{detail}}})\right)\right] \end{aligned} \tag{17}$$

其中 D_{detail} 表示细节层判别器，$D_{\text{detail}}(\cdot)$ 表示细节层判别器对输入图像真假的判断值；$I_{\text{detail-max}}$ 表示多光谱图像细节层与全色图像细节层的进行最大值处理得到的图像。

本文算法所设计的损失函数针对多光谱与全色图像各自的特点，主要从包含丰富光谱信息的基础层与包含清晰纹理信息的细节层两个层面进行设计，从而从主体和细节两个层面对融合图像保留的信息进行约束。

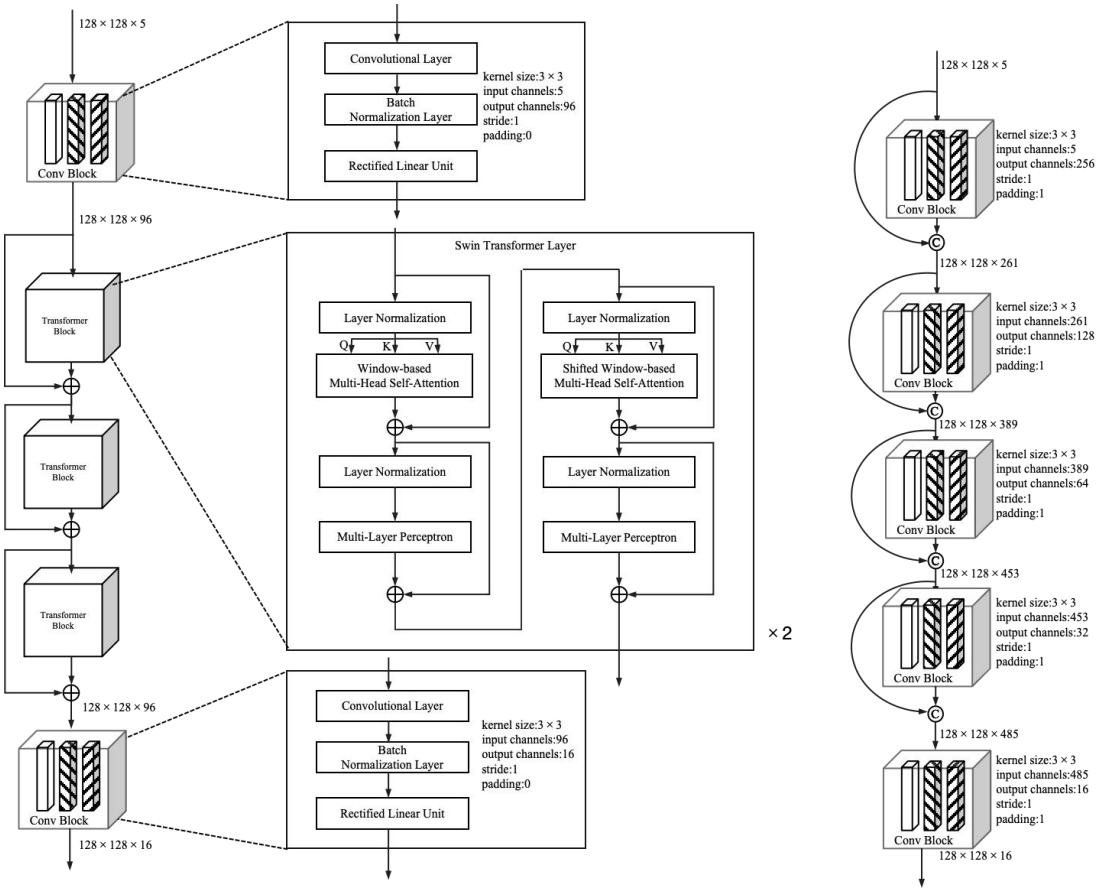
2.3 网络结构

CNN 具有强大的局部特征提取能力, 本文在生成器的细节层支路中采用卷积神经

网络，以更好地实现对细节层图像中清晰的空间细节信息进行提取。Transformer 具有较好的全局信息提取能力，本文在生成器的基础层支路中采用 Transformer 网络，以更好地实现对基础层图像中丰富的光谱信息进行提取。在判别器网络结构的设计中，本文均采用了 CNN 的网络结构。

2.3.1 生成器网络结构

本文所提方法中，生成器网络包括两条支路，即基础层与细节层两条支路。本文在设计生成器的网络结构时，针对基础层图像和细节层图像各自不同的特点，分别为其设计了不同的网络结构。如图 2(a)所示，基础层网络由两个 Conv Block 和三个 Transformer Block 组成。第一个 Conv Block 用来初步提取浅层信息，其输入为包括 4 个多光谱通道与 1 个全色通道的基础层图像。卷积核大小设置为 3×3，步长为 1，卷积核个数为 96。得到的特征图将被输入到三个结构相同的 Transformer Block 中。本文在多个 Transformer Block 之间采用残差连接，以提升网络的收敛速度、增强网络的表达能力并改善梯度消失的问题。Transformer Block 采用比 ViT 结构计算时间更短、效率更高的 Swin Transformer（Liu 等，2021）网络结构。每个 Transformer Block 由两个相同的 Swin Transformer Layer 组成，其中 Swin Transformer Layer 包含两个多层感知机 MLP（Multi-Layer Perceptron）、一个基于窗口的多头自注意力机制 W-MSA（Window-based Multi-Head Self-Attention）与一个基于移动窗口的多头自注意力机制 SW-MSA（Shifted Window-based Multi-Head Self-Attention），并在每个多头自注意力机制与每个多层感知机前添加一个层归一化 LN（Layer Normalization），在每个模块后采用残差连接。



(a) 生成器基础层网络结构 (b) 生成器细节层网络结构

(a) Network structure of generator basic layer (b) Network structure of generator detail layer

图 2 生成器网络结构图

Fig.2 Network structure of generator

每个 Transformer Block 的输入为形状固定的 128×128×96 的特征图，特征图首先被 8×8 的局部窗口分割成 256 个 8×8×96 的特征图，然后对每个窗口特征图 $\boldsymbol{F}_{window}$ 分别做多头自注意力计算，其中计算过程中的查询特征矩阵 $\boldsymbol{Q}$（Query）、键特征矩阵 $\boldsymbol{K}$（Key）和值特征矩阵 $\boldsymbol{V}$（Value）表示为：

$$\begin{aligned}\boldsymbol{Q}&=\boldsymbol{F}_{window}\cdot\boldsymbol{M}_{Q}\\ \boldsymbol{K}&=\boldsymbol{F}_{window}\cdot\boldsymbol{M}_{K}\\ \boldsymbol{V}&=\boldsymbol{F}_{window}\cdot\boldsymbol{M}_{V}\end{aligned}\tag{18}$$

其中 $\boldsymbol{M}_{Q}$、$\boldsymbol{M}_{K}$、$\boldsymbol{M}_{V}$ 为投影矩阵。通过自注意力机制计算得到局部窗口的注意力矩阵，其计算过程表示为：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{S}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\boldsymbol{T}}}{\sqrt{\boldsymbol{d}}}+\boldsymbol{p}\right)\cdot\boldsymbol{V}\tag{19}$$

其中 $S(\cdot)$ 表示归一化指数函数（Softmax）；$\boldsymbol{d}$ 表示维度；$\boldsymbol{p}$ 表示可学习的相对位置编码。接着将多头自注意力输出的注意力矩阵送入层归一化中，然后送入到多层感知机中进行位置编码与特征映射的非线性变换，最

终得到具有全局特征的特征图。

每个 Transformer Block 的输出均为 128×128×96 的特征图，将最后一个 Transformer Block 送入一个用以改变特征图维度的 Conv Block 中, 其卷积核大小为 3×3, 步长为 1，卷积核个数为 16。最后输出大小为 128×128×16 的特征图，即基础层支路的输出特征图。

生成器细节层支路主要提取细节层图像中的清晰纹理信息，故采用卷积神经网络进行特征提取，如图 2(b)所示。每个卷积层之间采用了残差连接，可以较好地将浅层信息传递到深层网络中，使网络更高效。每个卷积块均采用了批处理归一化以克服对数据初始化的敏感性，从而较好地避免梯度爆炸问题。卷积块的激活函数均采用 ReLU 函数以增强网络稳定性，加快收敛速度。细节层支路与基础层支路的输出特征图大小相同，将两个支路输出的特征图相加后送入卷积核大小为 3×3、步长为 1、卷积核个数为 4、激活函数为 Tanh 的图像重建层，继而得到生成器最终的输出图像，即融合图像。

2.3.2 判别器网络结构

判别器分为基础层判别器与细节层判别器，两个判别器的网络结构相同，均如图 3 所示：

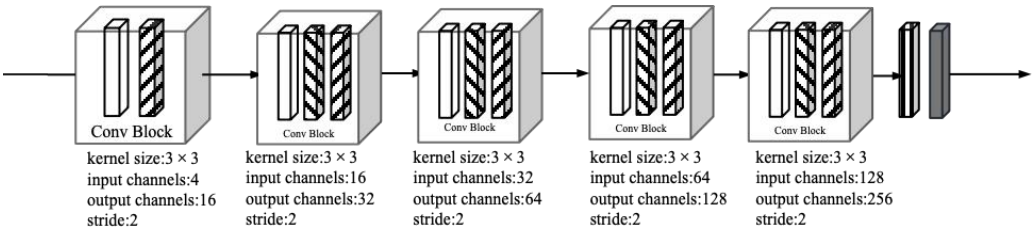


图 3 判别器网络结构图

Fig.3 Network structure of discriminator

其中每个卷积核大小均为 3×3，输入通道数分别为 4、16、32、64、128，步长均为 2。五个卷积层后是一个全连接层与激活函数 Tanh。基础层判别器的输入为源多光谱图像基础层与源全色图像基础层经平均值化处理后的基础层图像，以及融合图像的基础层；细节层判别器的输入为源多光谱图像细节层与源全色图像细节层经最大值化处理后的细节层图像，以及融合图像的细节层。

3 实验结果与分析

本文所提模型由一个生成器与两个判别器组成，其中两个判别器分别为基础层判别器和细节层判别器。具体地，针对基础层图像和细节层图像各自的特性，即基础层图像具有丰富的光谱信息、细节层图像具有清晰的纹理信息等，本文在生成器中设计了两条不同的支路，实现对基础层图像和细节层图像分别进行处理。此外，将融合图像进行分解，以便从基础层与细节层两个方面对模型进行对抗训练和约束优化。为了验证所提方法模型的性能，本文选取九种基于不同方法的先进模型，从主观与客观两方面对不同方法的融合结果进行定性与定量的对比分析。所有的对比方法均根据相关的原始文献进行参数设置。此外，为了证明模型中对图像分层处理的有效性，本文进行了消融实验。所有实验通过云计算完成，云计算的具体配置为：CPU 参数为 12 核、磁盘内存 32G；GPU 采用显存 24G 的 GeForce RTX 3090。此外，实验均在 Pytorch 深度学习框架下完成。

3.1 数据集介绍

本文实验使用的数据来自于三个公开的遥感数据集 (Yang 等, 2022; Meng 等, 2020)，包括 WorldView II、IKONOS 和 Quick Bird 三种数据。WorldView II 数据集涵盖建筑物、植被、裸地、水体等地物信息，具有高度的空间分辨率，地物纹理细节清晰，其全色图像空间分辨率为 0.5m，多光谱图像空间分辨率为 2.0m；IKONOS 数据集具备较好地光谱还原，覆盖建筑物、阴影、植被，其全色图像空间分辨率为 1m，多光谱图像空间分辨率为 4m；Quick Bird 数据集涵盖城市建筑、农田、裸地、植被、水体，具有大量的以城市建筑物为主的地物信息，广泛应用于测绘制图、城市详细规划、环境管理、农业评估等领域，其全色图像空间分辨率为 0.61m，多光谱图像空间分辨率为 2.44m。三种数据集中的多光谱图像均包括红光波段、绿光波段、蓝光波段和近红外波段。本文中，将三个数据集的全色与多光谱图像分别各裁成 4800 组一一对应的图像块，总计 14400 组图像块，其中全色图像块大

小为 128×128，多光谱图像块为 32×32。

在本文实验中，从裁剪好的图像块中随机选取 13 组作为测试数据，其余图像块作为训练集，在预处理中，对多光谱图像进行四倍上采样，以保持多光谱图像与全色图像输入网络的尺寸一致。

3.2 参数设置

在本文实验中，网络的初始学习率设置为1×10^{-4}，batch size 设置为 4，epoch 设置为 10，用于网络训练的优化器为 RMSprop；生成器损失函数中，λ_1 设置为 100，λ_2 设置为 1，a 设置为 0.7，b 设置为 0.7，c 设置为 0.7，α 设置为 0.1，β 设置为 0.1，γ 设置为 0.2，ω 设置为 0.2。

3.3 对比算法

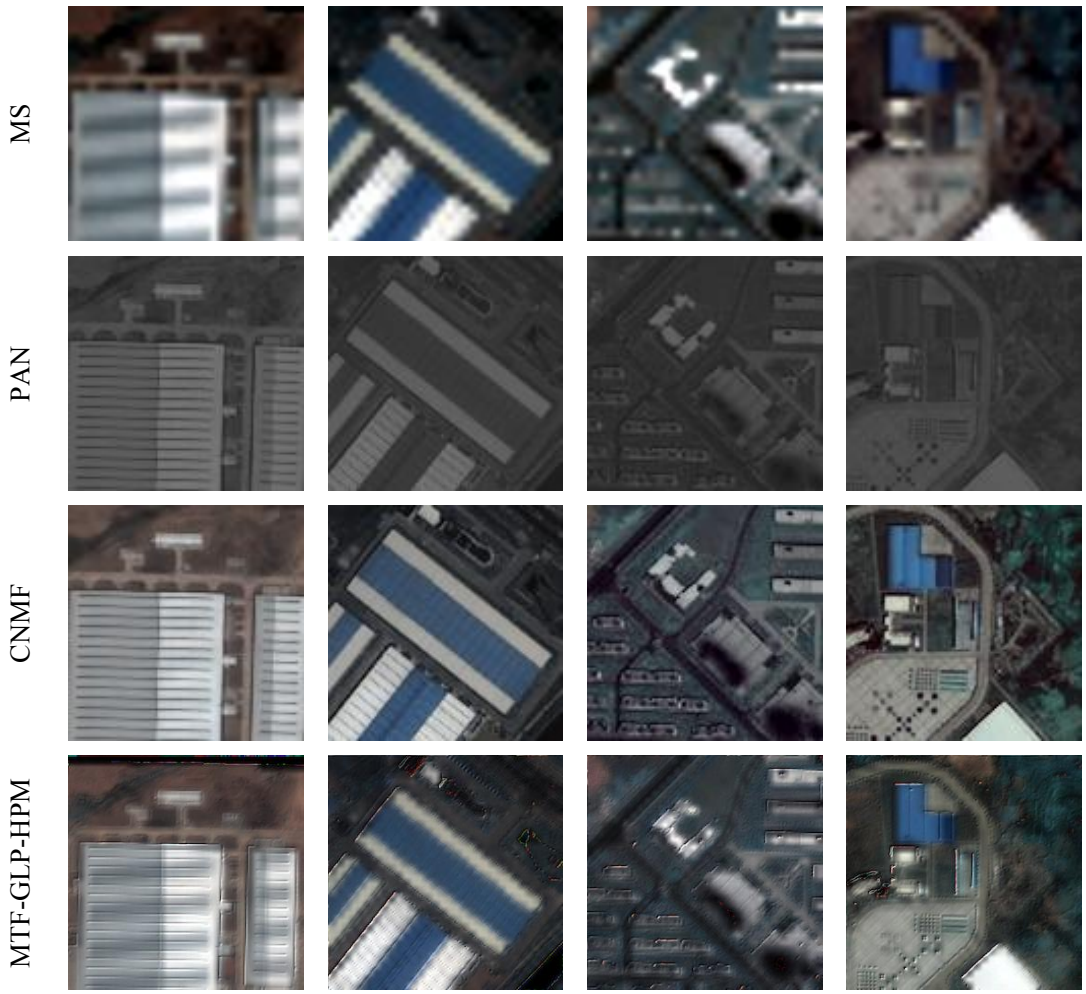
本文从预处理（裁剪）后的数据集中随机选择了 13 组一一对应的多光谱图像与全色图像作为测试数据，与九种具有代表性的先进算法进行定性与定量对比。本文中，用于对比实验的方法包括：CNMF（Yokoya 等，2011），MTF-GLP-HPM（Vivone 等，2013），SFIM (Liu , 2000)，GSA (Aiazzi 等，2007)，PanNet (Yang 等, 2017)，Pan-GAN (Ma 等, 2020)，UCGAN (Zhou 等, 2022)，SDPNet (Xu 等, 2020) 和 GTP-PNet (Zhang 等，2021）。其中，耦合非负矩阵分解方法 CNMF（Coupled Nonnegative Matrix Factorization）作为具有代表性与融合效果较好的传统方法，具有更新规则快捷、易于实现等特性，从而能够在空间域和光谱域生成较高质量的融合图像；MTF-GLP-HPM 是一种基于广义拉普拉斯金字塔的融合方法，通过调制传递函数匹配的滤波器将信息注入融合图中；SFIM 是一种基于亮度调节的平滑滤波方法，通过平滑滤波将全色图像与多光谱图像进行匹配，以得到融合图像；基于自适应 Gram-Schmidt 正交化方法 GSA，通过逆变换用全色图像替换多光谱图像的空间分量，以得到融合图像；PanNet 作为深度神经网络在遥感图像融合领域的早期代表性网络，具有较好地融合效果；Pan-GAN 作为首次基于生成对抗网络的无监督 MS 和 PAN 遥感图像融合方法，具有一定的代表性；UCGAN 作为基于周期一致性和生成对

抗网络的无监督MS和PAN遥感图像融合方法，具有较好的融合效果；SDPNet由两个编码器组成，分别用于提取遥感图像的浅层特征与深层特征，在基于编码器-解码器网络的遥感图像融合方法中具有较好的效果；GTP-PNet由梯度转换网络与残差网络组成，具有网络表达能力强、易于优化等优势，融合性能较强。综上，本文选取以上九种包括具有代表性与优异性能的传统方法、基于生成对抗网络的方法、基于卷积神经网络的方法以及基于深度残差网络的方法进行对比实验。

3.4 结果对比与分析

3.4.1 主观视觉效果对比

源多光谱图像MS、源全色图像PAN以及基于九种对比方法和本文提出方法的融合结果如图4所示。



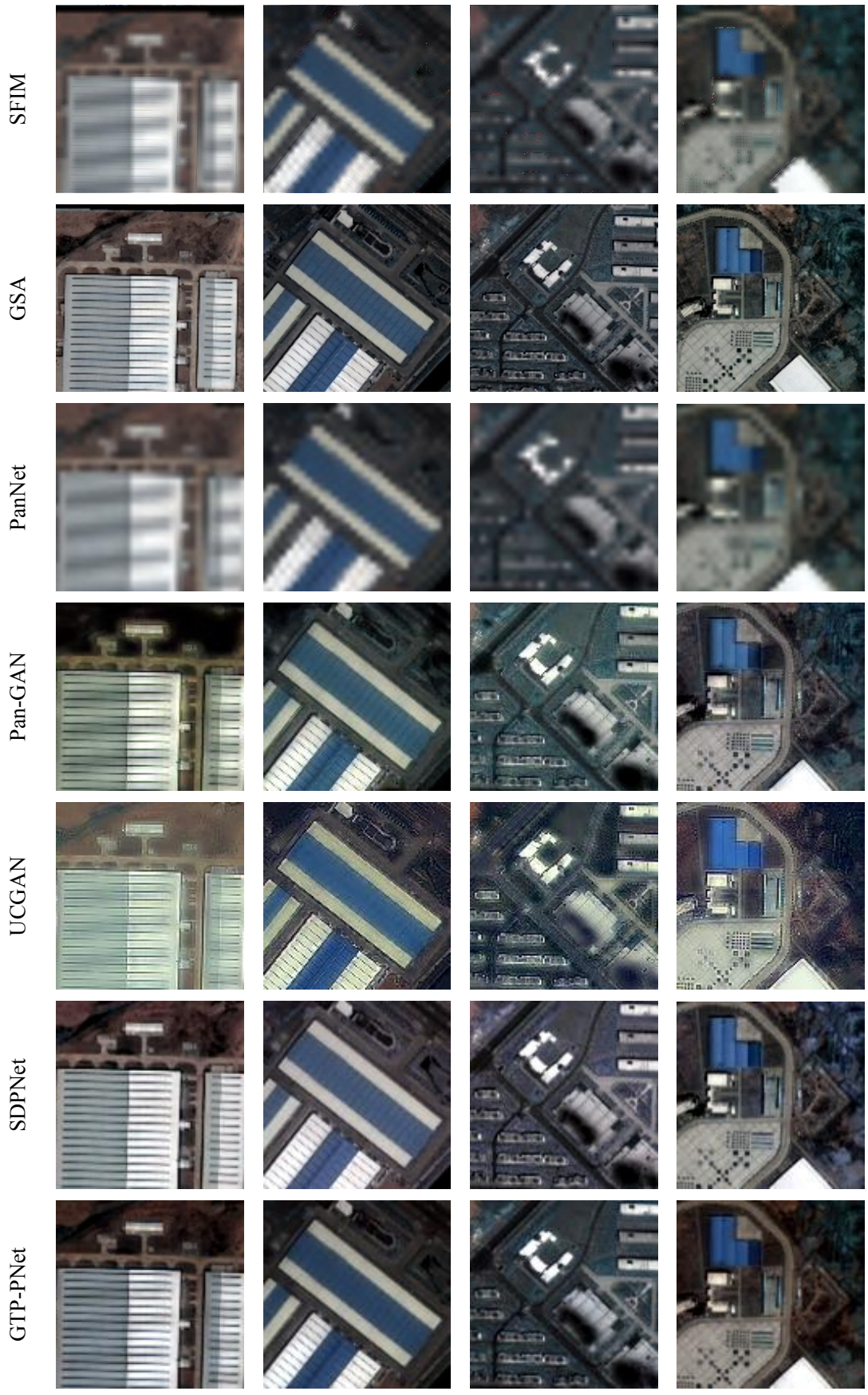

SFIM
GSA
PanNet
Pan-GAN
UCGAN
SDPNet
GTP-PNet

图 4 不同方法的融合结果对比

Fig.4 Comparison of fusion results using different methods

在图 4 所展示的四组融合结果中，CNMF 方法得到的融合结果纹理信息较好，但光谱信息存在一定程度的失真；MTF-GLP-HPM 方法与 SFIM 方法光谱信息保存较好，但纹理信息较为模糊；GSA 方法得到的融合结果视觉效果较好，纹理信息较为清晰，且光谱信息准确；PanNet 方法得到的融合结果的视觉效果较差，特别是空间纹理信息不足，如第一、二组图中房顶上的平行纹理过于淡化。Pan-GAN 方法得到的融合结果纹理信息清晰，光谱信息较好；UCGAN 方法得到的融合结果的光谱信息有些失真，如在第四幅融合图像中，存在假彩色现象。此外，UCGAN 方法得到的融合结果纹理信息较模糊，如第一、二幅融合图像的屋顶线条信息不清晰。SDPNet 方法与 GTP-PNet 方法得到的融合结果光谱信息完整、空间纹理信息清晰，但一些纹理细节信息存在不同程度的模糊，如第二组图中房顶上的平行纹理淡化、第四组图中道路纹理模糊等。本文所提方法的融合结果图一方面既保留了源全色图像的清晰纹理信息；另一方面，与源多光谱图像相比，又在一定程度上避免了光谱失真。综上，相较于大多数对比方法，本文所提方法在整体上取得了较好的融合效果，即融合结果图既具有较好的空间纹理信息，又较好地保持了源多光谱图像的丰富光谱信息。

3.4.2 客观评价指标对比

为了进一步比较和评估所提算法的图像融合性能，本文采用了六种遥感图像融合中常见的、有代表性的客观评价指标，从光谱信息与空间信息两个角度对不同方法的融合性能进行比较。本文选用的评价指标包括：信息熵 IE（Information Entropy）、光谱角 SAM（Spectral Angle Mapper）、均方根

误差 RMSE（Root-Mean-Square Error）、通用图像质量指数 UIQI（Universal Image Quality Index）（Wang 等，2002）、光谱失真指数 D_λ（Alparone 等，2008）与峰值信噪比 PSNR（Peak Signal-to-Nosie Ration）。

IE 主要用以衡量融合图像包含信息的丰富程度，计算公式如式（20）所示，其中融合图像的灰度分布为 $p=\{p_0,p_1,\cdots,p_i,\cdots,p_{L-1}\}$，$p_i$ 表示第 i 个灰度级像素个数与总像素个数的比，IE 越大，表明融合图像的信息越丰富，图像质量越好；

$$IE=-\sum_{i=0}^{L-1}p_i\log_2 p_i \tag{20}$$

SAM 主要用以衡量融合图像与源多光谱图像之间的光谱扭曲程度，计算公式如式 (21) 所示，SAM 越小，表明融合图像的光谱失真越小，图像质量越好；

$$SAM=\arccos\left(\frac{\langle I_{\mathrm{MS}}\cdot I_{\mathrm{F}}\rangle}{\|I_{\mathrm{MS}}\|_2\|I_{\mathrm{F}}\|_2}\right) \tag{21}$$

RMSE 用以计算融合图像与源图像之间的差异，其计算公式如式（22）所示，RMSE 越小，融合图像质量越好；

$$RMSE=\sqrt{\frac{\sum_{x=1}^{H}\sum_{y=1}^{W}\left(I_{\mathrm{MS}}(x,y)\right)^2}{\left(-I_{\mathrm{F}}(x,y)\right)}}{H\times W} \tag{22}$$

UIQI 用以衡量融合图像结构的扭曲程度，如式（23）所示，其中 σ_1、σ_2 分别表示融合图像与多光谱图像的标准差，C 表示协方差，K 为常数，

$$UIQI=\frac{4\sigma_1\sigma_2 C}{\left(\sigma_1^{\,2}+\sigma_2^{\,2}\right)\left(C^2+K\right)} \tag{23}$$

UIQI 越大，表明融合图像结构扭曲程度越小，图像质量越好；D_λ 用以计算融合图像的光谱失真程度，如式（24）所示，D_λ 越小，表明融合图像的光谱失真越小。

$$D_\lambda=\sqrt[p]{\frac{\sum_{x=1}^{H}\sum_{y=1}^{W}\left|Q\left(\hat{I}_{\mathrm{MS}},\hat{I}_{\mathrm{F}}\right)-Q\left(\tilde{I}_{\mathrm{MS}},\tilde{I}_{\mathrm{F}}\right)\right|^p}{H\times W}} \tag{24}$$

其中 $\hat{I}_{\mathrm{MS}}$、$\hat{I}_{\mathrm{F}}$ 与 $\tilde{I}_{\mathrm{MS}}$、$\tilde{I}_{\mathrm{F}}$ 表示多光谱图像与融合图像经不同特征矩阵相乘得到的结果；$Q(\bullet)$ 的计算如下：

$$Q(x,y)=\frac{4\sigma_{xy}\cdot\bar{x}\cdot\bar{y}}{\left(\sigma_x^2+\sigma_y^2\right)\left(\bar{x}^2+\bar{y}^2\right)} \tag{25}$$

PSNR 用以评估图像的保真程度,计算公式如式（26）所示，其中 MAX 表示源图像与融合图像间取最大值，MSE 表示源图像与融合

图像的均方误差，PSNR 越大，融合图像质量越好。

$$PSNR = 10\log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{26}$$

采用 IE、UIQI、PSNR 指标从空间信息角度对不同方法的融合结果进行比较；采用 SAM、RMSE、D_λ 指标从光谱信息角度对不同方法的融合结果进行比较。在计算与光谱信息角度相关的评价指标前，因多光谱图像与融合图像的尺寸存在差异，因此，本文将多光谱图像进行四倍上采样操作，以保证多光谱图像与融合图像的尺寸一致。表 1 所示为不同方法取得的评价指标的平均值，其中加粗数值表示效果最优者，下划线表示效果次优者。在表 1 中，左半部分指标 (IE、UIQI、PSNR）即为从空间信息角度进行比较的客观指标、右半部分指标 (SAM、RMSE、D_λ) 即为从光谱信息角度进行比较的客观指标。

表 1 不同方法在融合 13 组图像时取得的评价指标的平均值

Table 1 Average value of evaluation metrics regarding 13 pairs of images fused by the different methods

Method Category	IE↑ 空间	UIQI↑ 空间	PSNR↑ 空间	SAM↓ 光谱	RMSE↓ 光谱	D_λ ↓ 光谱
CNMF	6.8827	0.4979	17.2454	1.5277	9.8401	0.0375
MTF-GLP-HPM	6.2020	0.7672	17.7682	1.5663	9.4210	0.0267
SFIM	6.1468	0.7846	17.9712	1.5662	9.4173	0.0277
GSA	6.3401	0.7228	17.3088	1.5659	9.4565	0.0421
PanNet	<u>6.9403</u>	0.5966	17.3362	1.5234	9.8767	0.0336
Pan-GAN	6.6907	0.5606	18.6934	1.5165	9.7231	0.0315
UCGAN	6.8852	0.5118	17.0403	<u>1.5118</u>	10.0309	0.0294
SDPNet	6.3745	0.7594	17.7837	1.5515	9.4800	0.0318
GTP-PNet	6.2446	**0.9017**	<u>20.1033</u>	1.5248	**8.8922**	<u>0.0266</u>
Ours	**7.1245**	<u>0.8415</u>	**22.7357**	**1.5030**	<u>9.1279</u>	**0.0229**

注：表中加粗字体为该指标中效果最优的结果，加下划线为该指标中效果次优的结果。

由表 1 可知，本文所提方法的融合结果在 IE、SAM、D_λ 、PSNR 指标上均为最优，在 RMSE、UIQI 指标上仅次于 GTP-PNet 方法，表明所提方法取得了总体最优的融合效果。需要指出的是，原全色图像的 IE 平均值为 5.5921，表明所有算法的融合图像中包含的信息量均得到了增加。为了更直观地展示不同方法在融合 13 组图像时取得的客观评价指标的详细情况，图 5 和图 6 分别以折线图和箱式图的形式展示了不同方法在六种

客观评价指标上的具体差异。由图 5 和图 6 可知，本文提出方法的融合图像具有更丰富的光谱信息与更清晰的空间纹理信息。

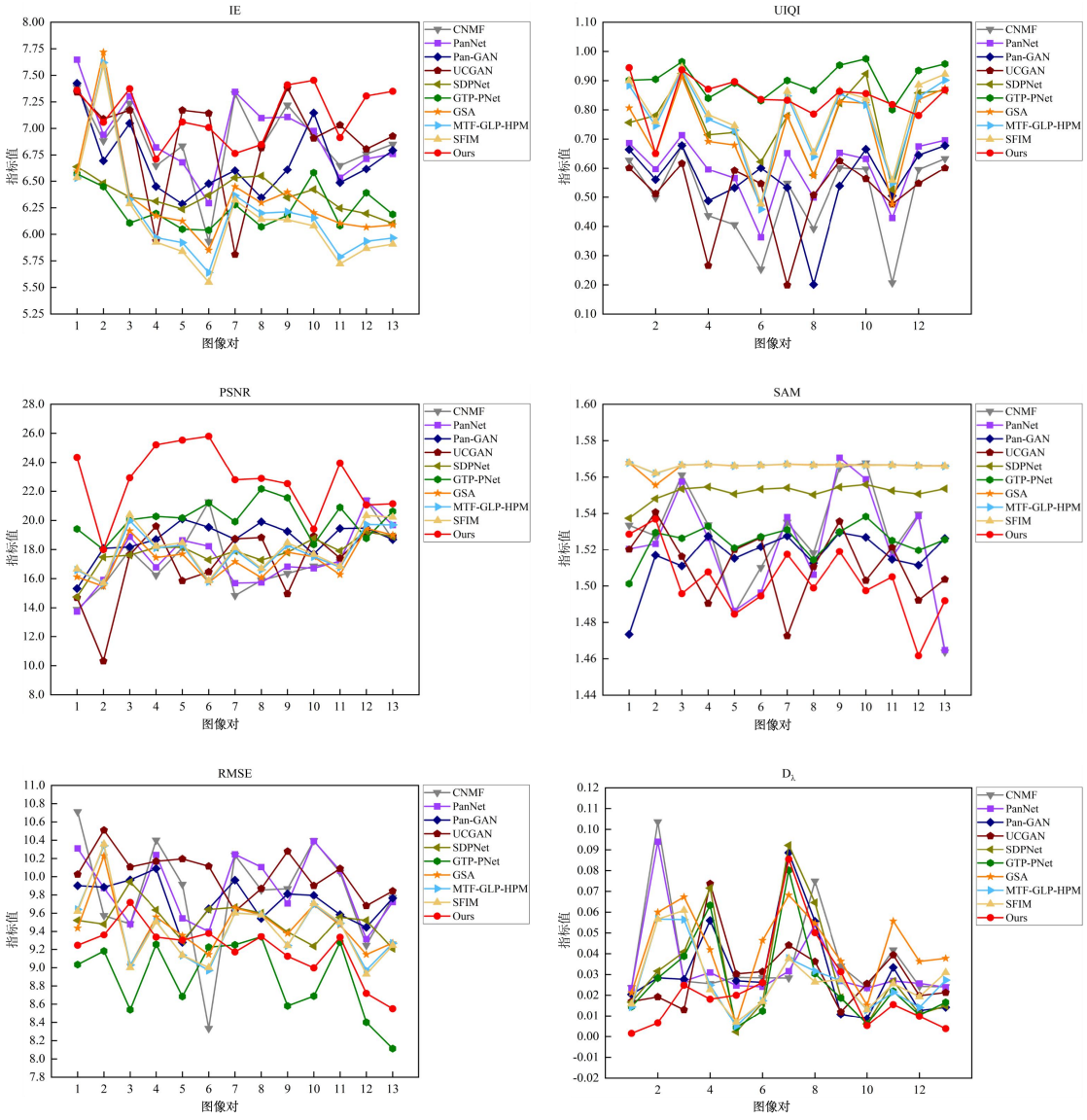
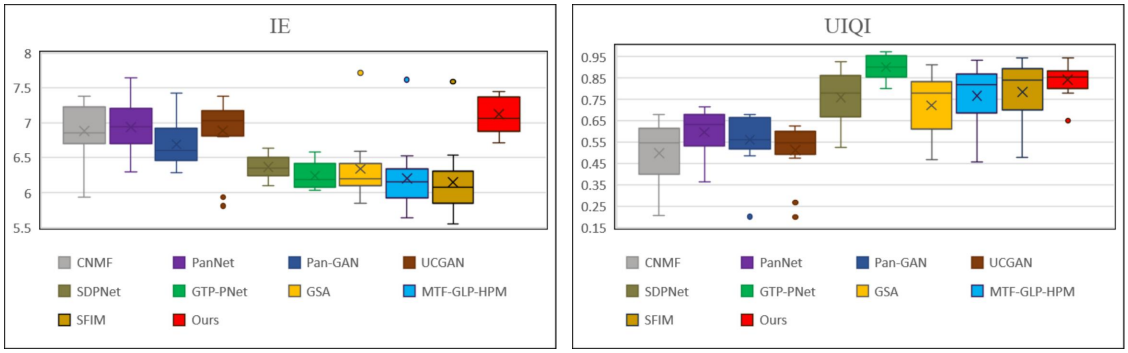


图 5 不同方法在融合 13 组图像时取得的评价指标的具体情况（折线图）

Fig.5 Detailed values of evaluation metrics regarding 13 pairs of images fused by the different methods (Line chart)



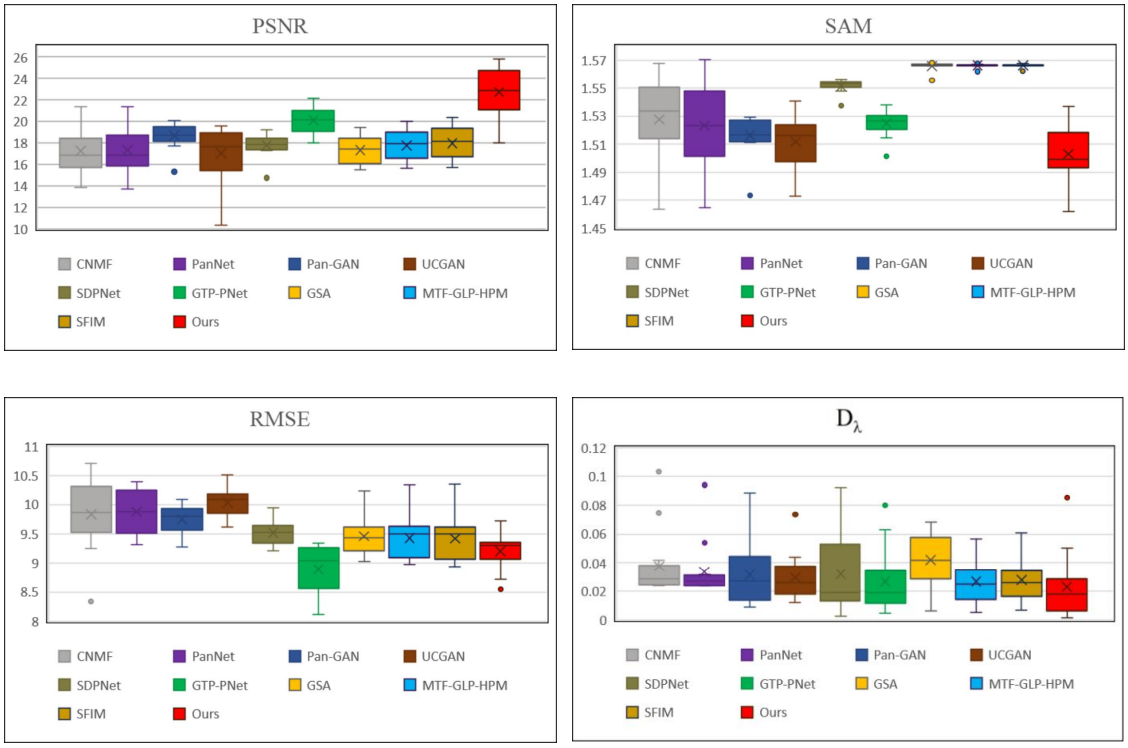
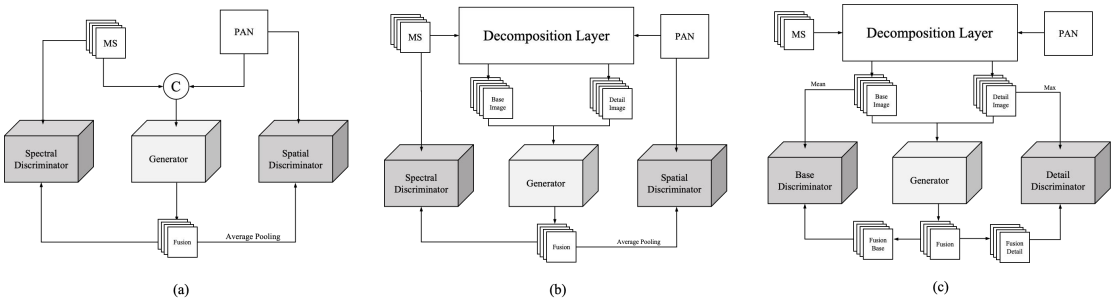


图 6 不同方法在融合 13 组图像时取得的评价指标的具体情况（箱式图）

Fig.6 Detailed values of evaluation metrics regarding 13 pairs of images fused by the different methods（Box chart)

3.4.3 消融实验

本节将模型框架结构划分为三种类型来进行消融实验和对比分析，三类模型分别是不进行图像分解的模型，仅对生成器输入进行图像分解的模型，以及对生成器的输入与输出均进行图像分解的模型，即本文所提模型。通过三类模型的融合实验来验证本文的图像分解模型及生成器双支路网络结构的优越性和有效性。三种模型的框架如图 7 所示。图 7(a)为不进行图像分解的模型(记为模型 I)，即将源多光谱图像与源全色图像在通道维度连接后输入至生成器；将生成器输出的融合图像输入到光谱判别器；将生成器输出的融合图像在通道维度上进行平均池化操作，使其通道数与源全色图像一致，然后将平均池化后的融合图像输入到空间判别器中。



(a) 无图像分解的模型 I

(a) The model I without image decompostion

(b) 仅对输入图像进行分解的模型 II

(b) The model II with decomposition only for inputted image

(c) 本文提出的模型

(c) The proposed model

图 7 三种模型的框架结构

Fig.7 Framework structure of three types of models

图 7(b)为仅对生成器输入进行图像分解的模型(记为模型 II)，即将源多光谱图像与源全色图像输入到分解层并得到各自的基础层图像与细节层图像，其中分解层的结构与图 1 融合框架图中提出的分解层相同；将基础层图像与细节层图像同时分别输入到生成器的两条支路中；将生成器输出的融合图像和经平均池化后的融合图像分别输入到光谱判别器与空间判别器。图 7(c)为本文提出的模型，该模型与图 7(b)的区别在于两个判别器不同，以及两个判别器的输入也不同。两种消融模型及本文所提模型的融合结果如图 8 所示。

(a) MS

(b) PAN

(c) Using ablation Model I

(d) Using ablation Model II

(e) Using ours

图 8 三种模型的融合结果

Fig.8 Fusion results of three models

图 8 中，从左到右分别为源多光谱图像、源全色图像、基于 7(a)所示消融模型的融合结果、基于图 7(b)所示消融模型的融合结果和基于图 7(c)所示的本文提出模型的融合结果。如图 8 所示，相较于消融模型(a)与消融模型(b)的融合结果，本文所提模型的融合结

果在主观上具有光谱信息完整、纹理细节清晰的优势。三种模型融合结果的客观指标如表 2 所示, 相较于两个消融模型的融合结果, 本文所提模型的融合结果在多个指标上均有较大的提升。综上，通过主观视觉效果与客观评价指标两方面的比较，证实了本文所提模型在尽可能保留更多源全色图像空间纹理信息的同时，较好地避免了光谱失真，保留了完整的光谱信息。

表 2 三种模型在融合 13 组图像时取得的评价指标的平均值

Table 2 Average value of evaluation metrics regarding 13 pairs of images fused by three models

Method	IE↑	UIQI↑	PSNR↑	SAM↓	RMSE↓	D_λ ↓
Category	空间	空间	空间	光谱	光谱	光谱
Ablation Model (a)	6.5277	0.5194	19.1657	1.5221	9.9958	0.0394
Ablation Model (b)	6.4754	0.5114	18.4848	1.5252	10.0406	0.0274
Ours	**7.1245**	**0.8415**	**22.7357**	**1.5030**	**9.1279**	**0.0229**

注：表中加粗字体为该指标中效果最好的结果。

4 结论

本文提出了一种无监督的基于双支路生成对抗网络与 Transformer 的全色和多光谱遥感图像融合方法。本文所提方法首先通过引导滤波将图像分解成基础层图像与细节层图像；接着，针对基础层图像光谱信息丰富、细节层图像空间纹理信息清晰的不同特点，分别采用全局特征提取能力较强的 Transformer 网络与局部细节特征提取能力较强的 CNN 提取特征。与此同时，所提方法模型中，设计了双判别器式生成对抗网络，包括具有并行特点的双支路网络结构的生成器，以及注重基础层光谱信息的基础层判别器和注重细节层空间纹理信息的细节层判别器。通过与多个模型进行定性与定量对比，验证了本文提出的用于全色和多光谱遥感图像融合模型的优越性。此外，通过消融实验的结果对比，进一步证实了本文所设计的网络结构的有效性。

虽然本文所提方法在融合全色和多光谱遥感图像时取得了一定的优势，但所提模型仍有一定的不足，如融合图像的空间纹理信息的清晰度尚有进一步提升的空间，部分原因在于仅采用引导滤波并不能完全将光谱信息与纹理信息分解。此外，本文所提方法虽对多光谱图像与全色图像之间的分辨

率差异无要求，但当多光谱图像分辨率与全色图像分辨率相差大于四倍时，本文方法的融合结果质量将存在一定程度的降低，其原因在于本文方法仅通过对多光谱图像多倍上采样，使源图像（多光谱图像与全色图像）保持一致的分辨率。因此，对于不同模态遥感图像融合的任务，如何进一步针对各自图像的不同特性进行分析研究，将是我们未来研究的重点，如改进现有的基于深度学习的融合方法，进一步深化不同深度学习网络之间的关系并有效结合，以实现不同深度学习网络间的优势互补；探索新型网络驱动的多光谱和全色遥感图像融合方法；探索遥感图像融合任务与其他遥感图像应用领域之间的结合，如遥感图像地物分类、遥感图像语义分割等。

参考文献（References）

Aiazzi B, Baronti S, Selva M. 2007. Improving component substitution pansharpening through multivariate regression of MS+Pan data. IEEE Transactions on Geoscience & Remote Sensing, 45(10): 3230-3239 [DOI:10.1109/TGRS.2007.901007]

Alparone L, Aiazzi B, Baronti S, Garzelli A, Nencini F and Selva M. 2008. Multispectral and panchromatic data fusion assessment without reference. Photogrammetric Engineering & Remote Sensing, 74(2): 193-200 [DOI: 10.14358/PERS.74.2.193]

Ballester C, Caselles V, Igual L, Verdera J and Rouge B. 2006. A variational model for P+XS image fusion. International Journal of Computer Vision, 69(1): 43-58 [DOI: 10.1007/s11263-006-6852-x]

Choi J, Yu K and Kim Y. 2010. A new adaptive component-substitution-based satellite image fusion by using partial replacement. IEEE Transactions on Geoscience and Remote Sensing, 49(1): 295-309 [DOI: 10.1109/TGRS.2010.2051674]

Deng L J, Feng M and Tai X C. 2018. The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-Laplacian prior. Information Fusion, 52: 76-89 [DOI: 10.1016/j.inffus.2018.11.014]

Dong Y, Cordonnier J B and Loukas A. 2021. Attention is not all you need: Pure attention loses rank doubly exponentially with depth//Proceedings of the 38th

International Conference on Machine Learning. Virtual: ACM: 2793-2803 [DOI: 10.48550/arXiv.2103.03404]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale//International Conference on Learning Representations. Virtual: Ithaca: 1-21 [DOI: 10.48550/arXiv.2103.03404]

Duran J, Buades A, Coll B, Sbert C and Blanchet G. 2017. A survey of pansharpening methods with a new band-decoupled variational model. ISPRS Journal of Photogrammetry and Remote Sensing, 125: 78-105 [DOI: 10.1016/j.isprsjprs.2016.12.013]

Garzelli A, Nencini F and Capobianco L. 2007. Optimal MMSE pan sharpening of very high resolution multispectral images. IEEE Transactions on Geoscience and Remote Sensing, 46(1): 228-236 [DOI: 10.1109/TGRS.2007.907604]

Han X H and Chen Y W. 2019. Deep residual network of spectral and spatial fusion for hyperspectral image super-resolution//IEEE Fifth International Conference on Multimedia Big Data (BigMM). Singapore: IEEE: 266-270 [DOI: 10.1109/BigMM.2019.00-13]

Imani M. 2019. Adaptive signal representation and multi-scale decomposition for panchromatic and multispectral image fusion. Future Generation Computer Systems, 99: 410-424 [DOI: 10.1016/j.future.2019.05.004]

Jin X, Huang S, Jiang Q, Lee S J and Yao S. 2020. Semi-supervised remote sensing image fusion using multi-scale conditional generative adversarial network with Siamese structure. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 7066-7084 [DOI: 10.1109/JSTARS.2021.3090958]

Li S T, Li C Y and Kang X D. 2021. Development status and future prospects of multi-source remote sensing image fusion. Journal of Remote Sensing, 25(1): 148-166 (李淑涛,李聪妤,康旭东. 2021. 多源遥感图像融合发展现状与未来展望. 遥感学报, 25(1): 148-166) [DOI: 10.11834/jrs.20210259]

Liu J G. 2000. Smoothing filter-based intensity modulation:

A spectral preserve image fusion technique for improving spatial details. International Journal of Remote Sensing, 21(18): 3461-3472 [DOI:10.1080/014311600750037499]

Liu Y, Chen X, Wang Z, Wang Z J, Ward, Rabab K and Wang X. 2018. Deep learning for pixel-level image fusion: Recent advances and future prospects. Information Fusion, 42: 158-173 [DOI: 10.1016/j.inffus.2017.10.007]

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S and Guo B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF international conference on computer vision. Virtual: IEEE: 10012-10022 [DOI: 10.48550/arXiv.2103.14030]

Ma F, Huo S and Yang F X. 2021. Graph-Based logarithmic low-rank tensor decomposition for the fusion of remotely sensed images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14: 11271-11286 [DOI: 10.1109/JSTARS.2021.3123466]

Ma J Y, Yu W, Chen C, Liang P W, Guo X J and Jiang J J. 2020. Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion. Information Fusion, 62: 110-120 [DOI: 10.1016/j.inffus.2020.04.006]

Ma J Y, Yu W, Liang P W, Li C and Jiang J J. 2019. FusionGAN: A generative adversarial network for infrared and visible image fusion. Information Fusion, 48: 11-26 [DOI: 10.1016/j.inffus.2018.09.004]

Ma L, Liu Y, Zhang X, Ye Y, Yin G and Johnson B A. 2019. Deep learning in remote sensing applications: A meta-analysis and review. ISPRS Journal of Photogrammetry and Remote Sensing, 152(6): 166-177 [DOI: 10.1016/j.isprsjprs.2019.04.015]

Meng X C, Xiong Y M, Shao F, Shen H F, Sun W W, Yang G, Yuan Q Q, Fu R and Zhang H Y. 2020. A large-scale benchmark data set for evaluating pansharpening performance: Overview and implementation. IEEE Geoscience and Remote Sensing Magazine, 9(1): 18-52 [DOI: 10.1109/MGRS.2020.2976696]

Pradhan P S, King R L, Younan N H and Holcomb D W. 2006. Estimation of the number of decomposition

levels for a wavelet-based multiresolution multisensor image fusion. IEEE Transactions on Geoscience and Remote Sensing, 44(12): 3674-3686 [DOI: 10.1109/TGRS.2006.881758]

Qiu K, Yi B S, Xiang M and Xiao Z. 2019. Fusion of hyperspectral and multispectral image by dual residual dense networks. Optical Engineering, 58(2): 1-8 [DOI: 10.1117/1.OE.58.2.023110]

Ranchin T, Aiazzi B, Alparone L, Baronti S and Wald L. 2003. Image fusion—The ARSIS concept and some successful implementation schemes. ISPRS Journal of Photogrammetry and Remote Sensing, 58(1-2): 4-18 [DOI: 10.1016/s0924-2716(03)00013-3]

Rao D Y, Xu T Y and Wu X J. 2023. TGfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. IEEE Transactions on Image Processing, 2023: 1-12 [DOI: 10.1109/TIP.2023.3273451]

Shao Z and Cai J. 2018. Remote sensing image fusion with deep convolutional neural network. Applied Earth Observations and Remote Sensing, 5(5): 1656-1669 [DOI: 10.1109/JSTARS.2018.2805923]

Tang W, He F Z and Liu Y. 2022. YDTR: infrared and visible image fusion via Y-shape dynamic transformer. IEEE Transactions on Multimedia, 2022: 1-16 [DOI: 10.1109/TMM.2022.3192661]

Thomas C, Ranchin T, Wald L and Chanussot J. 2008. Synthesis of multispectral images to high spatial resolution: a critical review of fusion methods based on remote sensing physics. IEEE Transactions on Geoscience and Remote Sensing, 65(4): 1301-1312 [DOI: 10.1109/TGRS.2007.912448]

Vivone G, Restaino R, Mura M, Licciardi G and Chanussot J. 2013. Contrast and error-based fusion schemes for multispectral image pansharpening. IEEE Geoscience and Remote Sensing Letters, 11(5): 930-934 [DOI: 10.1109/LGRS.2013.2281996]

Wang Z and Bovik A C. 2002. A universal image quality index. IEEE signal processing letters, 9(3): 81-84 [DOI: 10.1109/97.995823]

Xu H, Ma J Y, Shao Z F, Zhang H, Jiang J J and Guo X J. 2020. SDPNet: A deep network for pan-sharpening with enhanced information representation. IEEE Transactions on Geoscience and Remote Sensing,

59(5): 4120-4134 [DOI: 10.1109/TGRS.2020.3022482]

Yang J F, Fu X Y, Hu Y, Huang Y, Ding X H and Paisley J. 2017. PanNet: A deep network architecture for pan-sharpening. Proceedings of the IEEE international conference on computer vision, 2017: 5449-5457 [DOI: 10.1109/ICCV.2017.193]

Yang J X, Zhao Y Q and Jonathan C W. 2018. Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network. Remote Sensing, 10(5): 800-823 [DOI: 10.3390/rs10050800]

Yang S, Wang M, Jiao L, Wu R and Wang Z. 2010. Image fusion based on a new contourlet packet. Information Fusion, 11(2): 78-84 [DOI: 10.1016/j.inffus.2009.05.001]

Yang Y, Su Z, Huang S Y, Wan W G, Tu W and Lu H Y. 2022. Survey of deep-learning approaches for pixel-level pansharpening. Journal of Remote Sensing, 26(12): 2411-2432 (杨勇, 苏昭, 黄淑英, 万伟国, 涂伟, 卢航远. 2022. 基于深度学习的像素级全色图像锐化研究综述.遥感学报, 26(12): 2411-2432) [DOI: 10.11834/jrs.20211325]

Yokoya N, Yairi T and Iwasaki A. 2011. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. IEEE Transactions on Geoscience and Remote Sensing, 50(2): 528-537 [DOI: 10.48550/arXiv.2103.14030]

Zhang H and Ma J Y. 2021. GTP-PNet: A residual learning network based on gradient transformation prior for pansharpening. ISPRS Journal of Photogrammetry and Remote Sensing, 172: 223-239 [DOI: 10.1016/j.isprsjprs.2020.12.014]

Zhang K, Zhang F, Wan W B, Yu H, Sun J, Del S J, Elyan E and Hussain A. 2023. Panchromatic and multispectral image fusion for remote sensing and earth observation: Concepts, taxonomy, literature review, evaluation methodologies and challenges ahead. Information Fusion, 93: 227–242 [DOI: 10.1016/j.inffus.2022.12.026]

Zhou H Y, Liu Q J, Weng D W and Wang Y H. 2022. Unsupervised cycle-consistent generative adversarial networks for pan sharpening. IEEE Transactions on Geoscience and Remote Sensing, 60: 1-14 [DOI:

10.1109/TGRS.2022.3166528]

Zhou X R, Liu J, Liu S G, Cao L, Zhou Q M and Huang H W. 2014. A GIHS-based spectral preservation fusion method for remote sensing images using edge restored spectral modulation. ISPRS Journal of Photogrammetry and Remote Sensing, 88: 16-27 [DOI: 10.1016/j.isprsjprs.2013.11.011]

Panchromatic and multispectral remote sensing image fusion using dual-branch generative adversarial network combined with Transformer

Ji Yunxiang, Kang Jiayin, Ma Hanyan

School of Electronic Engineering, Jiangsu Ocean University, Lianyungang 222005, China

Abstract: Multispectral remote sensing image has rich spectral information that can reflect ground features, but its spatial resolution is low and its texture information is relatively insufficient. By contrast, panchromatic remote sensing image has high spatial resolution and rich texture information, but lacks rich spectral information that can reflect ground features. In practice, two kinds of images can be integrated into a single one to obtain the complementary advantages from the different images, thereby the fused image can better meet the needs of downstream tasks. To this end, this article proposes an unsupervised method for fusing the panchromatic and multispectral images using dual-branch generative adversarial network combined with Transformer.

Specifically, the source images (source panchromatic and multispectral images) are firstly decomposed into base and detail components using guided filtering, where the base component mainly focuses on the main body of the source image, and the detail component mainly represents the texture and detail information of the source image; Next, concatenates the decomposed base components of the panchromatic and multispectral images, and also concatenates the decomposed detail components of the two kinds of source images; Then, respectively inputs the concatenated base and detail components into the base and detail branches of the dual-branch generator; Next, according to the different characteristics of the base and detail components, respectively utilizes the Transformer and CNN to extract the global spectral information from the base branch and the local texture information from the detail branch; Then, continuously

trains the model in an adversarial manner between the generator and the dual discriminators (base layer discriminator and detail layer discriminator), and finally obtains the fused image with rich spectral information and high spatial resolution. Extensive experiments on the public dataset demonstrate that the proposed method outperforms the state-of-the-art methods both in qualitatively visual effects and in quantitatively evaluated metrics.

This article proposes an unsupervised fusion method for panchromatic and multispectral remote sensing images using dual branch generative adversarial network combined with Transformer. The superiority of the proposed method was verified via qualitative and quantitative comparisons with multiple representative methods. In addition, the ablation studies further confirm the effectiveness of the network structure designed in this article.

Key words: Remote sensing image fusion; Guided filtering; Convolutional neural network; Generate adversarial network; Transformer network; Basic layer; Detail layer; Panchromatic; Multispectral

Supported by National Natural Science Foundation of China (No. 62271236); Natural Science Foundation of Jiangsu Ocean University (No. Z2015009); Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX2022-41、No. KYCX2023-10)