

分类号: TH789

密 级: 公开

学校代码: 10363

学 号: 2210120135



安徽工程大学

Anhui Polytechnic University

硕士学位论文

题 目

基于改进自编码器的红外与

可见光图像融合算法研究

论 文 作 者

毛 飞

校 内 导 师

时培成（教授）

校 外 导 师

倪绍勇（高级工程师）

专业学位类别

机 械

研究方向（领域）

无人驾驶环境感知

论文提交日期: 2024 年 5 月 30 日

分类号：TH789

单位代码：10363

密 级：公开

学 号：2210120135

基于改进自编码器的红外与可见光图像融合算法研究
RESEARCH ON INFRARED AND VISIBLE IMAGE
FUSION METHOD BASED ON IMPROVED
AUTOENCODER

学生姓名： 毛飞

指导教师： 时培成（教授）

校外导师： 倪绍勇（高级工程师）

专 业： 机械

研究方向： 无人驾驶环境感知

论文答辩日期： 2024 年 5 月 26 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名： 

日期：2024年5月30日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

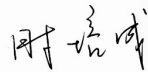
保密 ☐，在 年解密后适用本版权书。

本学位论文属于

不保密 ☐。

学位论文作者签名： 

日期：2024年5月30日

指导教师签名： 

日期：2024年5月30日

基于改进自编码器的红外与可见光图像融合算法研究

摘要

图像融合在智能驾驶中扮演关键角色，通过整合来自不同传感器的图像信息，从而创造信息更丰富的环境视图。作为图像融合领域的一个重要分支，红外与可见光图像融合旨在将红外图像的突出目标和全局信息与可见光图像的高清晰度和丰富纹理相结合，从而增强智能驾驶系统对周围环境的理解力。传统的图像融合方法存在计算复杂度高、模型调整困难等局限，近年来，自编码器网络凭借其出色的特征提取与图像重构能力在图像融合领域展现出了显著优势。基于此，本文提出了两种基于改进自编码器的红外与可见光图像融合算法，旨在进一步提升融合效果。主要研究内容如下：

（1）针对现有融合算法常从单一尺度提取特征，导致融合图像特征表现不全面、不均衡等问题，提出了一种基于双尺度编码结构与注意力机制的融合算法。该算法由编码器、融合策略和解码器三个部分组成。首先，在使用八度卷积块对源图像进行双尺度划分的基础上，采用密集链接模块和语义信息模块构建特征提取网络；接着，在编码网络的末端加入注意力机制 CoTNet 捕捉全局信息的依赖关系，强化重点信息表达；然后，把通道注意力机制与空间注意力机制接入融合策略，防止重要特征的丢失；最后，融合后的特征被传递进解码网络并重组成信息丰富的融合图像。该算法分别在 TNO

和 LLVIP 数据集上进行了融合测试实验，并与 7 种优秀的融合算法进行了比较。在主观肉眼判断和客观评价指标上都取得了优秀的结果，可以有效地进行红外与可见光图像的融合。

(2) 针对现有融合方法无法完全利用图像深层信息的问题，提出了一种基于多尺度编码结构与特征增强的融合算法 DAE-Nest。该算法由编码器-特征增强-融合层-解码器模块进行无监督学习。首先，利用多尺度编码子网络提取源图像特征；接着，将提取到的多尺度特征进行特征增强，以完全利用深层的图像信息；然后，将增强后的特征按照相应的融合策略融合；最后，解码器得到融合特征的输入并将其重组成富含丰富信息的融合图像。将 DAE-Nest 分别在 TNO 和 Road_Sence 数据集上进行相应的融合测试实验，并与 9 种优秀的融合算法进行比较。融合结果表明所提算法在主客观上都取得了优秀的结果，相比于其他代表性融合算法具有更优异的图像融合能力，能够为后续的视觉任务奠定良好的基础。

关键词：图像融合，红外与可见光图像融合，深度学习，自编码器，多尺度特征

RESEARCH ON INFRARED AND VISIBLE IMAGE FUSION METHOD BASED ON IMPROVED AUTOENCODER

ABSTRACT

Image fusion plays a crucial role in autonomous driving by integrating image information from different sensors to create a more information-rich environmental view. As an important branch in the field of image fusion, infrared and visible image fusion aims to combine the prominent targets and global information of infrared images with the high definition and rich texture of visible images, enhancing the autonomous driving system's understanding of the surrounding environment. Traditional image fusion methods are limited by high computational complexity and difficulties in model adjustment. In recent years, autoencoder networks have shown significant advantages in the field of image fusion due to their excellent feature extraction and image reconstruction capabilities. Based on this, this paper proposes two infrared and visible image fusion algorithms based on improved autoencoders, aiming to further enhance fusion effects. The main research contents are as follows:

(1) Addressing the problem of existing fusion algorithms often extracting features from a single scale, resulting in incomplete and unbalanced feature representation of the fusion image, a fusion algorithm based on dual-scale encoding structure and attention mechanism is proposed. This algorithm consists of three parts: encoder, fusion strategy,

and decoder. Firstly, on the basis of dividing the source image into dual scales using octave convolution blocks, a feature extraction network is constructed using dense connection modules and semantic information modules. Secondly, an attention mechanism CoTNet is introduced at the end of the encoding network to capture the dependency relationship of global information and enhance the representation of key information. Then, the channel attention mechanism and spatial attention mechanism are introduced into the fusion strategy to prevent the loss of important features. Finally, the fused features are passed to the decoding network and reorganized into an information-rich fused image. This algorithm was tested on the TNO and LLVIP datasets and compared with 7 excellent fusion algorithms. It achieved excellent results in subjective visual judgment and objective evaluation metrics, effectively performing infrared and visible light image fusion.

(2) Addressing the problem that existing fusion methods cannot fully utilize deep image information, a fusion algorithm DAE-Nest based on multi-scale encoding structure and feature enhancement is proposed. This algorithm performs unsupervised learning using an encoder-feature enhancement-fusion layer-decoder module. Firstly, the multi-scale encoding sub-network is used to extract features from the source image. Then, the extracted multi-scale features are enhanced to fully utilize the deep image information. Next, the enhanced features are fused according to the corresponding fusion strategy. Finally, the decoder receives the input of the fused features and reorganizes them into a fused image rich in information. DAE-Nest was tested on the TNO and Road_Sense datasets and compared with 9 excellent fusion algorithms. The fusion results show that the proposed algorithm achieved excellent results both subjectively

and objectively, compared to other representative fusion algorithms, demonstrating superior image fusion capabilities and laying a good foundation for subsequent visual tasks.

KEY WORDS: image fusion algorithm, infrared and visible image fusion, deep learning, autoencoder, multi-scale features

目 录

摘 要.....	I
ABSTRACT.....	III
第 1 章 绪论.....	1
1.1 研究背景与意义.....	1
1.2 国内外研究现状.....	2
1.2.1 基于传统的融合方法.....	3
1.2.2 基于深度学习的融合方法.....	4
1.2.3 现有融合方法的局限性.....	9
1.3 本文的主要贡献与创新.....	10
1.3.1 主要研究内容.....	10
1.3.2 创新点.....	11
1.4 本文的具体研究内容与结构安排.....	12
1.5 本章小结.....	13
第 2 章 红外与可见光图像融合研究理论基础.....	14
2.1 红外与可见光成像原理及特点.....	14
2.1.1 红外图像成像原理及特点.....	14
2.1.2 可见光图像成像原理及特点.....	14
2.1.3 红外图像与可见光图像优缺点.....	14
2.2 图像融合的层级分类.....	15
2.3 深度学习技术.....	17
2.3.1 深度学习概述.....	17
2.3.2 卷积神经网络.....	17
2.3.3 残差网络.....	21
2.3.4 注意力机制.....	22
2.4 自编码器及其变体网络.....	22
2.4.1 自编码器介绍.....	22

2.4.2 自编码器网络结构.....	23
2.4.3 其它变体网络.....	24
2.5 融合图像质量评价.....	25
2.5.1 主观评价方式.....	25
2.5.2 客观评价方式.....	25
2.6 本章小结.....	28
第 3 章 基于双尺度编码结构与注意力机制的融合算法.....	29
3.1 引言.....	29
3.2 融合网络介绍.....	29
3.2.1 八度卷积.....	29
3.2.2 注意力机制 CoTNet.....	30
3.2.3 主体网络结构.....	31
3.2.4 融合策略.....	33
3.2.5 训练阶段.....	34
3.3 融合实验及效果评估.....	36
3.3.1 实验环境及评价指标.....	36
3.3.2 参数选择.....	37
3.3.3 实验结果分析.....	38
3.3.4 消融实验.....	43
3.4 本章小结.....	43
第 4 章 基于多尺度编码结构与特征增强模块的融合算法.....	45
4.1 引言.....	45
4.2 融合网络介绍.....	45
4.2.1 网络结构.....	45
4.2.2 特征增强模块.....	47
4.2.3 融合策略.....	48
4.2.4 训练策略.....	50
4.2.5 损失函数.....	52
4.3 融合实验及效果评估.....	53

4.3.1 实验环境及评价指标.....	53
4.3.2 参数选择.....	54
4.3.3 实验结果分析.....	58
4.3.4 融合结果检测实验.....	63
4.3.5 消融实验.....	64
4.3.6 网络融合效率分析.....	65
4.4 DSAFusion 和 DAE-Nest 算法对比	66
4.5 本章小结.....	67
第 5 章 总结与展望.....	69
5.1 总结.....	69
5.2 展望.....	70
参考文献.....	71
攻读硕士学位期间发表学术论文及参研项目.....	78
致谢.....	79

插图清单

图 1-1 多源传感器信息融合	1
图 1-2 同一成像场景下的图像融合示例	2
图 1-3 红外与可见光融合算法综述	3
图 1-4 基于 VGG-19 的图像融合过程.....	5
图 1-5 基于 FusionGAN 的图像融合过程	6
图 1-6 基于 DenseNet 的图像融合过程.....	8
图 1-7 本文的章节结构	13
图 2-1 不同的融合层级结构图	16
图 2-2 LeNet 网络结构	17
图 2-3 简单的二维卷积	18
图 2-4 最大池化与平均池化计算过程	18
图 2-5 残差结构	21
图 2-6 SENet 注意力机制	22
图 2-7 编码器的基本结构	22
图 3-1 八度卷积	30
图 3-2 注意力机制 CoTNet.....	31
图 3-3 DSAFusion 算法结构	31
图 3-4 密集链接网络结构	32
图 3-5 语义信息网络结构	33
图 3-6 DSAFusion 融合网络结构	34
图 3-7 训练阶段网络结构	34
图 3-8 灰度转换示例	35
图 3-9 灰度转换后的数据集图例	36
图 3-10 不同算法在 TNO 测试集上的融合结果一	39
图 3-11 不同算法在 TNO 测试集上的融合结果二.....	39
图 3-12 不同算法在 LLVIP 测试集上的融合结果一.....	40

图 3-13 不同算法在 LLVIP 测试集上的融合结果二.....	41
图 4-1 DAE-Nest 网络结构	46
图 4-2 特征增强模块 En 网络结构	47
图 4-3 DAE-Nest 端到端的融合策略	49
图 4-4 AE 训练过程	50
图 4-5 融合策略训练过程	51
图 4-6 整体网络训练过程	53
图 4-7 灰度转换后的数据集图例	53
图 4-8 网络训练损失	55
图 4-9 不同 λ 取值下的融合实验结果.....	55
图 4-10 不同算法在 TNO 测试集上的融合结果一	59
图 4-11 不同算法在 TNO 测试集上的融合结果二.....	59
图 4-12 不同算法在 RoadScene 测试集上的融合结果一.....	61
图 4-13 不同算法在 RoadScene 测试集上的融合结果二.....	61
图 4-14 单目标场景下 YOLOv5 检测结果	63
图 4-15 多目标场景下 YOLOv5 检测结果	64

插表清单

表 3-1 DSAFusion 具体网络参数	36
表 3-2 不同参数网络融合结果评价指标平均值	37
表 3-3 不同算法在 TNO 测试集试验结果评价指标平均值	42
表 3-4 不同算法在 LLVIP 测试集试验结果评价指标的平均值.....	42
表 3-5 双尺度模块与注意力机制的消融实验	43
表 3-6 不同算法对 40 个图像对的平均处理时间.....	43
表 4-1 特征增强模块网络参数	47
表 4-2 融合策略的网络参数	50
表 4-3 AE 训练阶段网络设置	51
表 4-4 不同平衡参数融合结果评价指标平均值	56
表 4-5 使用不同融合策略融合结果平均指标	57
表 4-6 “IRD”中不同卷积核大小的“Conv7”的实验结果.....	57
表 4-7 不同算法在 TNO 测试集上的评价指标平均值	62
表 4-8 不同算法在 RoadScene 测试集上的评价指标平均值.....	62
表 4-9 多目标场景下的检测结果	63
表 4-10 特征提取尺度和特征增强模块消融实验	65
表 4-11 不同算法对 40 个图像对的平均处理时间.....	66

第1章 绪论

1.1 研究背景与意义

伴随人工智能技术的快速发展，智能化与自动化技术已经在汽车工业的各个方面获得应用，汽车逐渐成为高科技产品的集大成者。这种智能化不仅依赖于“大脑”计算机的高效分析能力，更重要的是依赖于汽车“眼睛”各种传感器所收集到的数据。随着越来越多种类车用传感器的出现，汽车获取周围环境信息的途径变得更加多元化，收集到的信息量也日益增加。在这样一个信息爆炸的时代背景下，如何有效地提取和整合这些庞大且复杂的数据，成为了亟待解决的问题。为此，多源传感器信息融合技术^[1]应运而生，它将不同传感器收集到的信息进行整合，使得处理后的信息更加全面和具体。这种技术实现了不同模态间信息的优势互补^[2]，可以有效地从复杂的数据中提取出重要信息，极大地增强了汽车对周边环境的感知能力，提升驾驶的安全性。图 1-1 展示了多源传感器信息融合的构成。

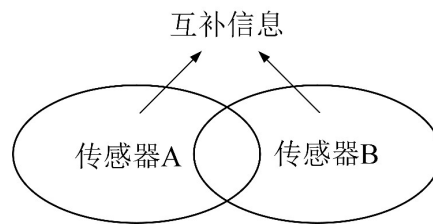


图 1-1 多源传感器信息融合

在汽车上的多种传感器收集到的信息种类中，视觉图像数据可以直观的传递给人们丰富的场景信息^[3]，但单一传感器的成像会受到成像原理、成像环境的限制，往往不能很好的反映出成像目标的场景信息，从而影响图像数据的分析效果。多源图像融合从属于多源传感器信息融合技术^[4]，本质^[5]是将包含不同信息的图像合并成单个具有增强可感知性的图像，旨在显著改善源图像。生成的融合图像相较于源图像包含更多可用的信息^[6]，有助于后续的视觉任务。多源图像融合不仅应用于自动驾驶环境感知，更在多种领域占据着重要地位，如物体检测^[7]、医学图像分析^[8]、遥感^[9]和军事应用^[10]等。

在多源图像融合领域，红外与可见光图像融合一直是研究热点^[11]。红外图像通过红外传感器捕捉物体发出的红外辐射生成，能够突出场景中的关键目标，

并具有在恶劣天气条件下（如雨、雪、雾霾）稳定成像的优势，可以提供丰富的热辐射和结构信息，但红外图像不能很好的表达细节纹理信息，且分辨率较低，不利于人的直观视觉感知^[12]。另一方面，可见光传感器通过反射光原理捕获自然光场景，呈现出与人眼直接观察相符的细节丰富的自然场景，具有高清晰度和符合人类视觉感知的特点^[13]。然而，可见光成像对环境变化较敏感，例如在雨雪、大雾天气或光线不足的条件下，成像效果会明显降低。因此，将两者进行融合，可以弥补彼此的缺陷，不仅可以提高对复杂场景的理解力，还有利于执行更精确的视觉任务。图 1-2 通过一个示例展示了在同一成像场景下红外与可见光图像融合，其中，红外图像提供场景下的显著目标信息，可见光图像提供丰富的细节纹理信息，二者相互补充生成一幅蕴含丰富信息的图像。

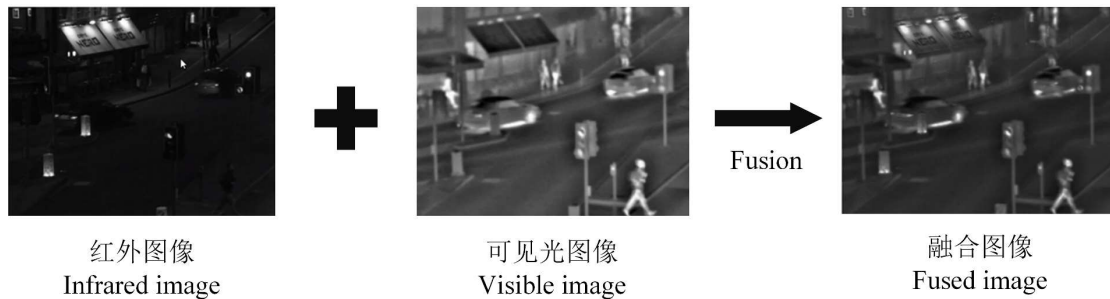


图 1-2 同一成像场景下的图像融合示例

红外图像与可见光图像因其各自独有的成像依据和光谱波段的差异，在诸多关键领域展现了显著的应用价值和研究潜力^[14]。红外图像能够在恶劣环境下突出重要目标，而可见光图像则提供丰富的场景细节，但二者信息的差异使得融合面临诸多挑战，如背景干扰、目标信息模糊以及细节信息损失等。鉴于此，本文提出两种基于改进自编码器的红外与可见光融合算法，旨在有效整合两种图像的优点，优化融合效果。

1.2 国内外研究现状

在红外与可见光图像融合研究的早期阶段，融合算法主要停留在像素级融合^[15]。但随着深度学习神经网络逐渐在图像处理方面展现出巨大的优势，图像融合技术得以迅速演进并拓展应用范围。现如今，可将红外与可见光图像融合技术分为基于传统的融合方法和基于深度学习的融合方法两类，本小节将从这两类融合方法进行详细的探讨，如图 1-3 所示。

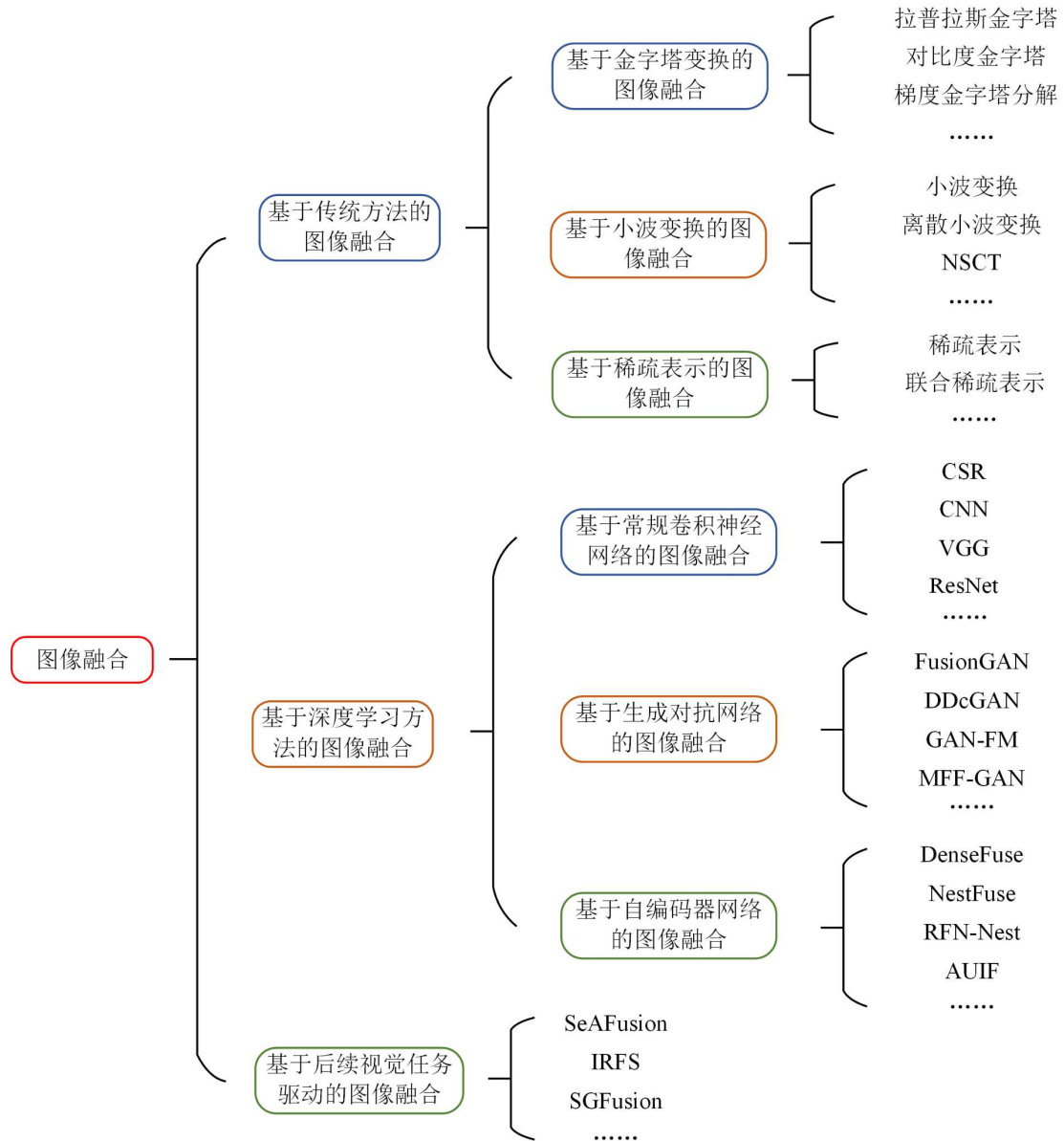


图 1-3 红外与可见光融合算法综述

1.2.1 基于传统的融合方法

传统的图像融合方法依据理论的不同，可分为基于金字塔变换的融合方法、基于小波变换的融合方法和基于稀疏表示的融合方法三类^[16]。

自 1980 年以来，基于金字塔变换的方法在图像融合领域取得了重大突破，这些方法通过分解、融合和重构三个主要步骤来完成图像的融合过程。1983 年，Burt 等人^[17]利用拉普拉斯金字塔对图像进行分解融合，标志着金字塔变换可以完成图像融合任务。1989 年，Toet 等人^[18]提出了基于对比度金字塔的融合方法，这种方法虽然能够有效地对不同模态的图像进行融合，但生成的图像往往伴随

有大量的噪声且表现出较低的融合质量。为了攻克这个难题，1993 年 Burt 等人^[19]利用梯度金字塔分解完成了图像融合。此后，像高斯金字塔^[20]、可控金字塔^[21]等方法也陆续被用来融合图像。虽然金字塔变换提供了有效的多尺度图像表示手段，但在融合过程中常导致重要信息的丢失，并且因为这些方法通常缺乏方向性，融合效果并不理想。

1990 年前后，小波变换技术的发展推动了图像融合领域的革新。基于小波变换的融合方法提供了一种在频域和时域同时进行分析的能力，允许对图像进行多尺度分解，能够保留更多的高频信息，使得其在细节保持方面表现更为出色。与金字塔变换相比，小波变换能够更精细地分析图像的频率信息，捕获图像的细节和纹理特征。基于小波变换的基础上，离散小波变换^[22]与双树复小波变换^[23]等技术也被应用于图像融合。然而，二维小波变换在处理图像时的各向同性特性限制了其在捕捉图像纹理、轮廓等信息上的能力。为了突破这个限制，Do 等人^[24]提出了基于轮廓波变换的方法，凭借更优的方向选择性和各向异性捕获能力，显著改善了小波变换在图像融合中的表现。随后，Yang 等人^[25]引入了基于非下采样轮廓波变换（NSCT）的技术，通过保持位移不变性，解决了轮廓波变换的局限性。这些技术的进展有效的优化了图像融合的效果。

基于稀疏表示的融合方法旨在通过优化图像的代表形式提高融合质量。这一过程首先需要将源图像进行分割；随后，对分割生成的每个小图像块进行稀疏编码以获得稀疏系数；接着通过预先设定好的融合规则，将来自不同源图像块的稀疏系数进行合并，这一步骤对决定最终融合效果十分重要。最后，融合后的稀疏系数将被同一过完备字典重新构建，从而生成融合图像。Yang 等人^[26]通过采用稀疏表示法成功实现了图像的融合，并展示了该方法的有效性。Tang 等人^[27]提出 MdedFusion，通过利用加权最小二乘法（WLS）进行图像分解，并使用局部统计边缘模型（LSEM）融合特征。尽管基于稀疏表示的融合方法在效果上具有显著优势，但其实现过程中需要构建高质量的过完备字典，这通常要求大量的计算资源和高质量的训练图像，此外，如何设计有效的融合规则也是一个关键问题，需要根据具体的应用场景和需求进行优化。

1.2.2 基于深度学习的融合方法

近期，在图像融合领域，深度学习技术的应用日益广泛。这种技术通过损

失函数引导的方式进行差异化的特征提取，实现了针对性更强的特征获取。这些方法主要针对特征提取、特征融合与图像重建进行优化。根据采用网络架构的不同，这些方法又可以分为基于卷积神经网络（CNN）、基于生成对抗网络（GAN）和基于自编码器（AE）三类。

2016年，Liu等人^[28]引入卷积稀疏表示（CSR）完成图像融合，有效地克服了传统稀疏表示技术在细节保留和图像误配准处理方面的限制。2017年Liu等人^[29]进一步提出了一种新的多焦点图像融合策略，利用深度卷积神经网络学习源图像与焦点图之间的直接映射，避免了传统方法在手工设计滤波器和规则比较清晰度信息时的复杂性。2018年Li等人^[30]将预训练的VGG-19网络应用于红外与可见光图像的融合，通过分解图像为基础和细节部分，并结合深度网络特征进行融合，具体的融合过程如图1-4所示。随后，ResNet也被用来融合图像^[31]，解决了直接使用深度特征可能导致的融合性能降低问题。上述融合方法都利用了大规模数据预训练的深度神经网络，以提取丰富的图像特征和有效的融合技巧，展现了优异的效率和准确性，为深度学习在图像融合领域的应用奠定了基础。然而，预训练模型在面对特定任务时可能会显示出一定的局限性，特别是在处理新颖或未见过的图像特征时。此时，生成对抗网络（GAN）^[32]提供了一个新的视角。

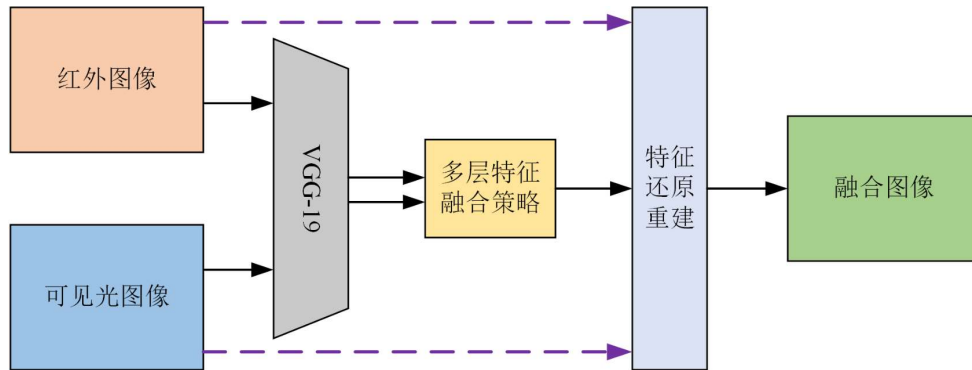


图 1-4 基于 VGG-19 的图像融合过程

2019年，Ma等人^[33]使用生成对抗网络（GAN）设计了名为FusionGAN的融合方法。该方法利用生成网络创造含有丰富红外图像信息的融合图像，并通过判别网络对融合图像与可见光图像进行区分。通过两者之间的对抗博弈，达到了融合的目的，该方法的融合过程如图1-5所示。然而，单个判别器难以有效地保留图像中的细节信息。随后Ma等人^[34]在2020年提出了一种采用双判别器

的改进 GAN 模型 DDcGAN，该模型通过使用两个判别器的策略，可以生成同时包含红外图像显著信息和可见光图像细节纹理的高质量融合图像。紧接着，2021 年，Zhang 等人^[35]开发了 GAN-FM，一个具有全尺寸跳跃连接的新型生成对抗网络，通过整合两个判别器，以更好地保留图像中的细节和纹理信息。2023 年，Liu 等人^[36]提出了一种基于注意力引导和小波约束的融合方法 AWFGAN，以解决现有方法在处理关键模态信息上的不足和双鉴别器之间的对抗问题。具体而言，该方法设计了两种独特的鉴别策略来提高融合性能：一是将空间注意力模块（SAM）引入生成器，通过空间注意力图强制使红外图像的鉴别集中于目标区域；二是将可见信息的鉴别范围扩展到小波子空间，从而强制生成器恢复可见光图像的高频细节。尽管上述方法通过对抗性训练显著提高了融合图像的质量，但 GAN 网络的训练依旧面临着高复杂度和低稳定性的挑战，特别是在保证模型收敛和生成图像一致性方面。

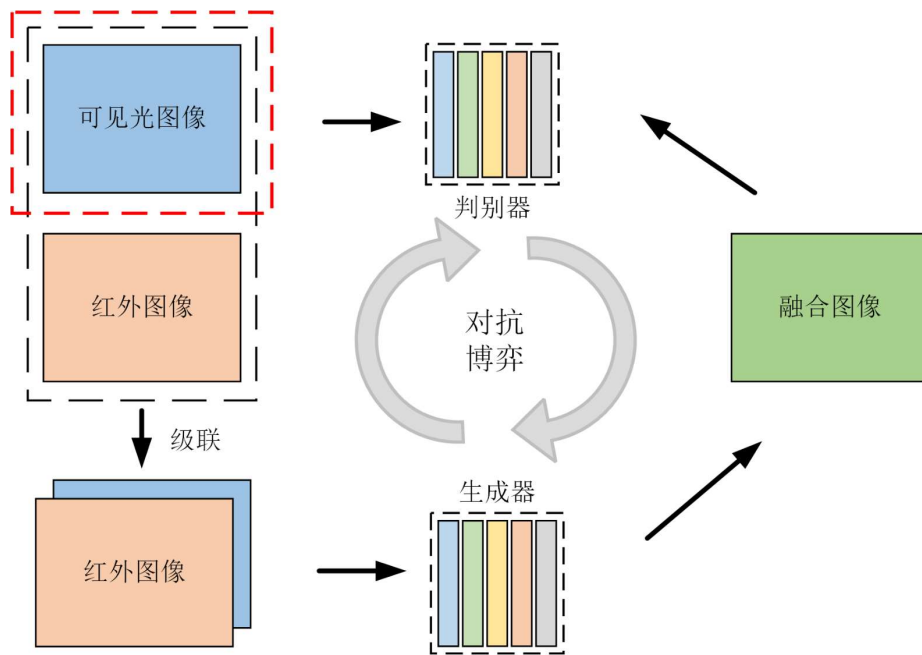


图 1-5 基于 FusionGAN 的图像融合过程

为了解决融合的稳定性问题并专注于学习数据的潜在表征，实现更加准确和精细的特征重建，自编码器（AE）^[37]也被应用于图像融合。2018 年，Li 等人^[38]设计了一种创新的图像融合框架 DenseFuse，该网络通过将密集块集成到编码器中，显著增强了从源图像中提取特征的能力，实现了特征的多次利用。该方法的核心优势在于它能够通过编码器精准地提取源图像的深层特征，保证了

图像中的关键信息被有效捕获。接着，根据既定的融合规则，将这些深层特征综合在一起，形成一个融合的特征集，包含了所有重要的信息。最后再通过解码器将这些融合特征重新映射回图像空间，完成最终融合图像的重建。此外，DenseFuse 还采用了结构相似损失和像素损失两种方式优化训练过程，以提升网络在图像重建任务上的表现，生成更高质量的融合图像。该融合方法的过程如图 1-6 所示。2020 年，Li 等人^[39]设计了一个基于嵌套连接形式的编码结构 NestFuse，该结构独特地采用了一种嵌套连接形式，这种设计允许它在不同的层次上有效地捕捉和整合源图像的信息，从而实现对图像多尺度特征的深入提取。随后，Li 等人^[40]在嵌套连接的基础上对融合网络进行了重新设计，提出了一种新的端到端的融合网络 RFN-Nest。该网络首先对红外与可见光图像进行下采样处理，以获得来自源图像的多尺度特征，随后，在残差网络的基础结构上对融合网络进行了全新设计，利用残差架构取代传统融合方法，并通过一个创新的“巢状”连接概念训练的自编码器，以及特别设计的保持细节和增强特征的损失函数来进行训练。2021 年，Zhao 等人^[41]提出了 AUIF，利用双编码器结构独立处理图像的基础层和细节层。这种方法使得 AUIF 能够精确地分离源图像中的低频成分（代表图像的基本形状和结构）和高频成分（负责捕捉边缘、纹理等细节信息），从而实现了针对不同尺度特征的有效提取。随后，通过对这些分解出的低频和高频特征进行专门的融合处理，AUIF 能够生成融合了所有重要视觉信息的图像，既保留了图像的基本内容，又增强了图像的细节表现。2022 年，Xue 等人^[42]提出的 FLFuse-Net 融合方法通过精确地从红外和可见光图像中提取共享及独有特征，创造出具有新视角的图像，不仅有效的增强了图像的视觉效果，还提高了信息的完整性和多样性。同年，Tang 等人^[43]提出了 PIAFusion，该网络通过自适应地维持显著目标的亮度分布，有效地保留了图像中的背景纹理信息，这一方法强调了在图像融合过程中对细节和背景纹理的保护，确保了最终融合图像在视觉上的自然度和真实感，同时增强了图像的表现力和丰富性。

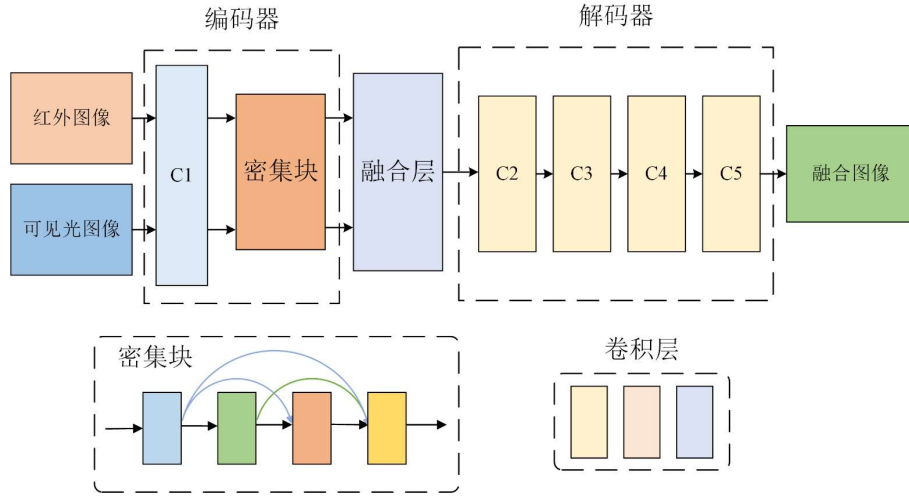


图 1-6 基于 DenseNet 的图像融合过程

此外，红外与可见光图像融合技术的发展也开始与其它视觉处理任务相结合，探索如何将图像融合技术与诸如目标检测、场景理解等其它视觉任务对齐，并设计出由这些后续任务驱动的融合方法，已经成为图像处理领域的一大研究趋势。2022 年，Tang 等人^[44]提出了一种创新的图像融合网络 SeAFusion，专注于通过融合技术提高图像的光照度，以便更有效地执行高级视觉任务。这种语义感知的实时图像融合方法利用深度学习技术理解图像内容的语义层面，旨在保留图像细节的同时，增强图像的亮度和对比度，从而使图像更适合于后续的视觉处理任务。Wang 等人^[45]在 2023 年将融合算法与显著目标检测算法集成，提出了一种称为 IRFS 的方法。该方法通过红外与可见光图像融合（IVIF）和多模态显著性目标检测（SOD）的结合，以发现和定位野外环境中难以发现的物体。通过特征筛选融合子网络（FSFNet）和融合引导的交叉补充 SOD 子网络（FC2Net）的设计，实现了模态间信息的有效融合和显著性特征的精确提取。同年，Liu 等人^[46]设计了 SGFusion，利用双引导编码、图像重建解码和显著性检测解码等过程，从源图像中提取特征和显著性信息，并利用显著性检测解码得到的显著性图作为融合权重，从而生成具有视觉感知性的融合图像。Zhou 等人^[47]利用人类视觉系统的特性，将图像特征转换到人类视觉响应空间，实现了在相同感知状态下融合所有特征，并采用双向显著性聚合和多尺度分解等技术，在消除外部物理因素干扰的前提下，实现了高质量的图像融合。这些跨任务的研究方法不仅提高了融合图像的实际应用价值，而且开拓了图像处理技术的新应用场景，有效的推动了视觉理解和分析技术的整体进步。

1.2.3 现有融合方法的局限性

国内外研究人员针对图像融合技术展开了广泛而深入的探索，但尽管传统方法和基于深度学习的方法都取得了一定的技术突破，仍面临着以下问题：

(1) 传统融合方法各有所长，但也存在诸多短板。基于金字塔变换的融合方法主要聚焦于空间尺度的分解，缺乏捕捉图像中方向性特征的能力，而且在对图像进行多层次分解的过程中，可能导致某些细节信息的丢失，特别是在图像重构阶段；基于小波变换的融合方法尽管比金字塔变换提供了更好的频率分解能力，但标准小波变换在捕捉图像的局部方向特征方面仍然有限，尤其是在高频细节的融合过程中，可能会在图像边缘引入伪影；而基于稀疏表示的方法能够灵活地适应不同的图像特征，但该方法的性能在很大程度上依赖于构建的过完备字典的质量，而构建一个能有效代表各种图像特征的字典是非常复杂和计算密集的，因为它不仅需要捕捉到复杂的图像特征，还要保持计算的可行性。总体来说，传统图像融合技术通常要求手动设计和调整图像变换以及融合规则，这在实际应用中可能会带来操作上的困难。此外，鉴于源图像本身的多样性和计算资源的限制，传统方法在处理大规模或高复杂度的图像数据时常显示出明显的局限性。

(2) 当前基于深度神经网络的每种方法都充分利用了深度学习的优势，取得了不错的效果，但也存在相应的不足。基于常规卷积神经网络的融合方法缺乏层次化的尺度分解，限制了其从源图像中提取全面和有效的多尺度特征的能力。此外，这种方法大量依赖预训练数据可能导致模型在特定应用场景下的适应性和泛化能力不足；基于生成对抗网络的融合方法虽然优化了图像生成的质量，却可能忽略了保留源图像中局部特征的重要性，限制了其在细节表现上的能力。加上 GAN 网络本身存在的稳定性和模型收敛问题，训练过程往往比较复杂；对于基于自编码器的融合方法，虽然在特定尺度上有效提取特征，却往往忽略了更深层次的图像信息，未能充分挖掘潜在的有用特征。此外，这种方法在特征融合策略上的不足，可能影响到最终融合效果的质量。总体而言，尽管基于深度学习的图像融合技术已显示出强大的潜力并取得了显著成果，它们各自存在的局限性也明确指出了未来研究的方向，即如何开发更高效的多尺度特征提取与融合策略，以及如何进一步提升模型的泛化能力。

1.3 本文的主要贡献与创新

1.3.1 主要研究内容

尽管目前的红外与可见光图像融合技术已经能够有效的执行融合任务，但仍存在上述局限性。将自编码器（AE）^[37]结构用于图像融合可以充分发挥其在特征提取及图像重建方面的优势，因此，本研究结合自编码器网络提出了两种改进的红外与可见光图像融合方法，旨在进一步改善图像融合的效果。具体的研究内容包括：

（1）总结相关基础理论知识，对现有融合算法展开广泛分析

深入分析和综合大量文献，聚焦红外与可见光图像融合展开调研，总结相关图像融合理论知识，广泛分析现有的传统方法和基于深度学习的方法，深入理解各种融合方法的优点与不足之处，针对不足之处展开研究，设计新型融合方案弥补缺点。

（2）研究基于双尺度编码网络与注意力机制的融合算法 DSAFusion

在源图像特征表征侧重不同的情况下，融合图像应取得适当的平衡。面对现有融合方法常从单一尺度提取特征的不足，结合密集链接模块、语义信息模块，设计一种双尺度编码网络，针对性的提取红外与可见光图像中不同尺度大小的特征信息。并在编码器网络末端加入注意力机制 CoTNet，同时利用通道注意力机制和空间注意力机制融合特征，聚焦重要信息的表达。

（3）研究基于多尺度编码网络与增强模块的融合算法 DAE-Nest

在图像融合过程中，采集浅层信息较为简单，深层信息的保留较为困难。针对深层图像特征较难保留的问题，提出基于多尺度特征提取的端到端的融合方法 DAE-Nest，首先，设计包含多尺度信息提取与特征增强模块的编码网络，对源图像包含的深层特征进行提取与增强。其次，根据图像特征中深浅特征表征的侧重不同问题，设计了新的融合模块 IRS（Inception ResNet Shallow）和 IRD（Inception ResNet Deep）来针对性的对不同层级的图像特征进行融合。最后，由基于巢连接形式的解码器对融合特征进行重建工作。

（4）利用主观视觉判断和客观评价指标对实验结果展开对比

本研究提出的两种红外与可见光图像融合方法均是基于自编码器网络框架，但 DAE-Nest 采用了端到端的融合策略，所以还需分步对其进行训练：阶段一训

练自编码器中编码器的图像编码能力以及解码器的特征还原重建能力；阶段二在阶段一的基础上进一步训练融合模块的特征融合能力。训练完成后，将所提两种算法在三个公开的常用红外与可见光图像数据集——TNO^[48]、LLVIP^[49]以及 RoadScene^[50]上进行了测试。通过与其它优秀融合方法进行比较分析，进一步证实了本研究所提融合方法的有效性。

1.3.2 创新点

基于本文主要的研究内容，总结以下创新点：

(1) 针对红外与可见光图像特性，考虑融合图像对不同源图像特征表征的平衡，本研究设计了一种新颖的双尺度卷积编码网络 DSAFusion。通过融入密集连接和语义信息模块，该网络针对性地提取不同尺度的特征信息，有效解决了单尺度特征提取的局限性。此外，编码网络末端集成了 CoTNet 注意力机制模块，以捕获全局依赖关系，使网络更集中于图像中的关键特征，避免了特征冗余。融合策略中还包括通道和空间注意力机制，优化特征融合步骤，保证重要特征表达。同时，在编码器、解码器中采用八度卷积代替标准卷积，以减少网络内存消耗，提高融合效率。

(2) 针对融合图像的成像质量，考虑充分提取与保留来自源图像更深层的信息，本研究提出一种端到端的融合网络框架DAE-Nest，从更深层的尺度上对源图像进行信息提取，并在对图像进行多尺度分解的基础上，将相对深层的特征信息加权到相对浅层中。这样不仅使除最深层外每一层的特征都得到了不同程度的增强，还使编码器提取到的特征在层级之间产生了信息交互。由于不同层级网络提取的特征存在差异，DAE-Nest中设计了新的端到端的融合模块IRS (Inception ResNet Shallow) 和IRD (Inception ResNet Deep)。其中，IRS适用于浅层尺度，可捕捉更多的细节、像素点信息，从而更好的提取与输入源相近的特征；IRD用于深层尺度则可获取图像中整体性较强的一些信息，进而得到语义性强，目标显著的特征。DAE-Nest利用改进的自编码器提取图像特征，并将解码器按照巢连接 (Nest) 的形式进行拼接。

1.4 本文的具体研究内容与结构安排

本论文的工作主要考虑如何改进自编码器网络框架以实现更好的红外与可见光图像融合效果，共分为5个章节，具体安排如下：

第1章：绪论。介绍红外与可见光图像融合的研究背景与意义，综述国内外研究者在此领域内的研究情况，并对本研究的主要工作和章节安排进行了简单介绍。

第2章：红外与可见光图像融合研究理论基础。首先介绍红外与可见光图像的成像原理以及独有特点；然后介绍图像融合的层级分类；接着介绍有关深度学习以及自编码器的基本理论；最后介绍本研究实验所采用的融合图像质量评估方法。

第3章：基于双尺度编码网络与注意力机制的红外与可见光图像融合算法DSAFusion。首先阐述了所提模型的设计思想；然后详细介绍双尺度编码网络与加入的注意力机制的模型细节；最后通过一系列对比实验论证本章融合算法的有效性。

第4章：基于多尺度编码网络与特征增强模块的红外与可见光图像融合算法DAE-Nest。首先对算法的设计动机进行了探讨；随后详细介绍了算法的结构；最后通过系列消融实验以及对比试验进行了广泛的分析，验证所提融合算法的有效性。

第5章：总结与展望，对本研究进行的主要工作进行总结并探讨了未来的研究趋势，为未来工作的开展提供了思路。

论文的章节结构图如图1-7所示。

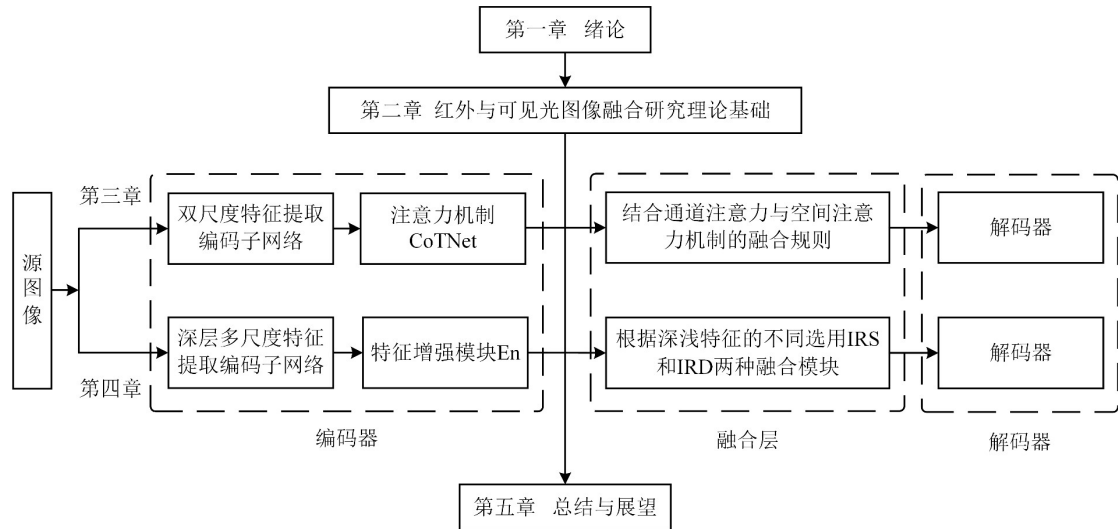


图 1-7 本文的章节结构

1.5 本章小结

本章首先对图像融合技术的研究背景进行了概述，接着介绍了当前领域的研究进展和现状，并在总结现有研究成果的基础上，分析了当前融合方法面临的主要挑战和问题，明确本研究的目标以及创新点。最后，给出了全文的章节结构安排。

第2章 红外与可见光图像融合研究理论基础

本文所介绍的两种红外与可见光图像融合算法均是基于自编码器（AE）网络改进而来。自编码器是一种使用深度学习技术来学习数据表示的人工神经网络模型。因此，本章节首先简要回顾了红外图像和可见光图像成像的基本原理及其主要特点，随后详细讲述了深度学习、自编码器网络和注意力机制的基础知识。最后，本章还介绍了用于评估融合图像质量的方法，这些内容为接下来的研究部分提供了坚实的理论基础。

2.1 红外与可见光成像原理及特点

2.1.1 红外图像成像原理及特点

红外线是由 0.78 到 1000 μm 波长的电磁波组成，由从近红外到远红外的各个波段一起构成了红外热辐射^[12]。在自然界中，所有物体均发出红外辐射，而红外传感器则捕捉这些辐射并将其转换为电信号，最终生成红外图像^[51]。这种成像技术能够通过测量目标与背景之间的温差来进行成像，特别适合于恶劣的视觉环境，如烟雾、雨雾中，因为它很少受外界环境因素影响，具有较强的探测能力。但是红外图像常常分辨率较低，缺少详细的图像细节，这在信息处理和描述方面带来了挑战。

2.1.2 可见光图像成像原理及特点

可见光传感器捕获自然场景的反射光^[13]，生成能还原出物体色彩的可见光图像。作为日常生活中常见的光源，可见光的波长范围介于 0.38 至 0.76 μm 。这种传感器产生的图像紧贴人眼的视觉习惯，来源包括但不限于手机照片和摄像机视频等。相比红外图像，可见光图像在标准环境下提供更高的分辨率和更直观的场景，更符合人类的视觉体验。不过，这类图像容易受到光照、时间及拍摄条件的影响，在低光或视线受阻的环境下，其成像质量会大幅下降。

2.1.3 红外图像与可见光图像优缺点

红外图像与可见光图像各自依据不同的成像特性展现了独特的优势和局限，具体分析如下：

(1) 成像环境需求：红外图像依赖于物体发射的红外辐射进行成像，对环境条件的要求较低；相比之下，可见光图像需要充足的光照和稳定的成像条件，对成像环境的要求相对较高。

(2) 成像质量差异：在光照充足的条件下，红外图像捕捉的信息量依然较少，对细节的表征能力不足；可见光图像则信息丰富、清晰度高。然而，在雨、雪、雾霾等恶劣天气情况下，红外图像的性能受影响较小，能提供比可见光图像更多的信息。

(3) 分辨率对比：由于工作波段和成像环境的限制，红外图像的分辨率通常较低；而可见光图像，得益于其与自然光线的关联及符合人眼观察的特性，具有更高的分辨率。

(4) 对比度分析：红外图像在展现目标与背景的对比度上更为突出，易于区分；而可见光图像的对比度信息较弱，区分度不如红外图像。

结合上述分析，红外图像与可见光图像各展所长，二者表现出显著的互补性。通过应用图像融合技术，不仅可以克服各自的局限和单一传感器成像的不足，还能有效地减轻成像环境的制约，提升了融合图像的全面性和实用性。

2.2 图像融合的层级分类

图像融合技术根据处理图像信息深度的不同，可分为像素级^[15]、特征级^[52]和决策级^[53]融合。三者从简单到复杂，各自针对图像处理的不同阶段，提供独特的融合策略。

(1) 像素级融合

像素级融合是图像融合中最基础的方法，通过直接操作源图像的像素值来融合，目的是尽可能地保留原始信息。然而，这一方法要求对源图像进行精确的配准，融合预处理阶段需要处理大量信息，进而增加程序的运行时间和处理的复杂度。此外，由于直接操作像素，该方法的融合效果容易受到噪声的影响，可能会降低最终图像的质量。

(2) 特征级融合

特征级融合先提取图像的特征，这一层次不直接处理原始像素，而是在图像的特征表示（如边缘、角点、纹理等）层面进行信息的整合，其目标是在融合图像中保留源图像的关键特征，如边缘和对比度，以此来提升融合图像的整

体质量。与直接在像素级进行融合的方法相比，特征级融合通过减少数据处理量的同时，依旧保持了图像的重要属性，使其更适用于复杂的分析任务，并在图像融合领域得到了广泛的应用。然而，因为特征级融合仅关注特定的图像特征而非整体像素信息，可能会导致一定量的信息丢失。

（3）决策级融合

决策级融合是在图像处理的最高层次上进行，即在做出某些判决或分类后，综合不同来源的决策结果。决策级融合在计算效率和容错性上具有显著优势，能够有效消除依赖单一传感器可能引入的误差。尽管决策级融合能够通过汇总多源决策信息来提升决策的准确性和可靠性，但其需要依赖于更底层的像素级和特征级融合结果。往往在整合过程中会丢失大量细节信息，在真正应用中使用较少，一般仅适用于特定的任务。

这三个层级的融合方法各具特点，选择合适的融合层级可以根据具体应用需求和预期目标来决定。如图2-1所示的结构图，这三种层次并非完全独立，有时会结合使用。因此，在处理具体的图像融合任务时，需要根据应用场景进行细致分析，选择合适的融合层次以达到最佳效果。

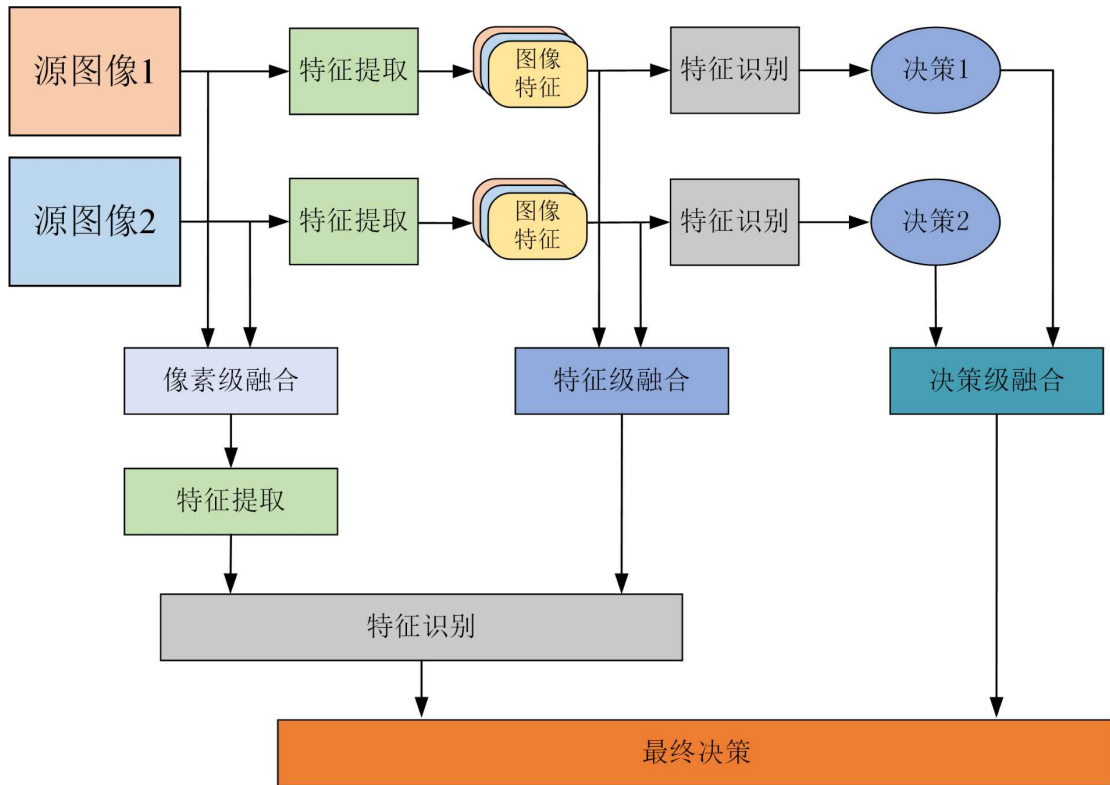


图 2-1 不同的融合层级结构图

2.3 深度学习技术

2.3.1 深度学习概述

近年来，深度学习在图像处理和模式识别领域应用逐渐宽泛^[54]，起源于对人工神经网络的研究^[55]，强调利用多隐层构建模型。旨在通过深度训练识别数据中的关键特征，Hinton等^[56]提出的深度信念网络展示了其在特征提取方面的能力。不同于传统方法的人工特征选择，深度学习自动学习数据中的复杂特征，揭示内在规律，用于分类、识别等任务。它通过逐层提取并构建复杂特征，解决识别问题。随着技术进步和数据集增多，深度学习成为多领域的关键技术。

2.3.2 卷积神经网络

卷积神经网络（CNN）^[57]是深度学习中最常用的神经网络模型之一，以其能通过训练学习高级特征和简化的结构著称。CNN 通过局部连接和权值共享减少了模型参数，简化了训练过程^[58]。LeNet^[59]作为首个 CNN 模型在手写数字识别任务中，展现了卓越的性能。LeNet 的网络结构如图 2-2 所示，通过两层卷积和池化后，利用全连接层输出分类结果。LeNet 的成功促进了 CNN 在图像处理等领域的广泛应用，为 CNN 模型的进一步发展奠定了基础。

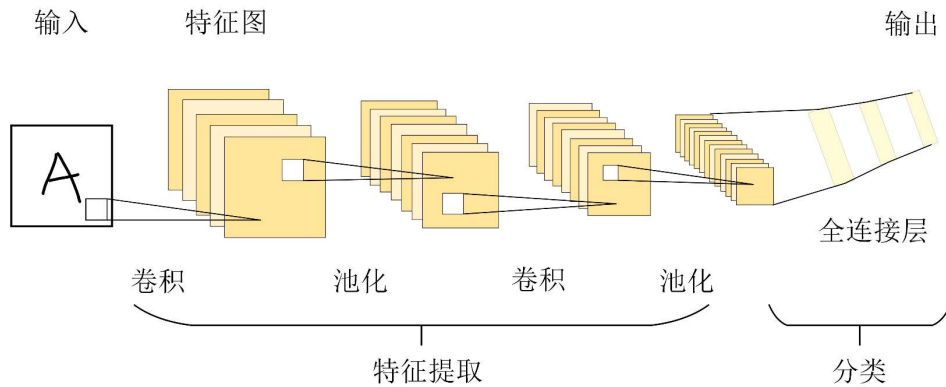


图 2-2 LeNet 网络结构

2.3.2.1 卷积层

卷积层从输入的图像或前一层特征图中进行图像特征的提取，该过程通过卷积核在特征图上按照设定的步长执行卷积操作，从而捕捉局部区域的特征，生成目标特征图。此操作有效地捕获图像中的细节和结构信息，为深层次特征提取和图像识别任务提供基础。卷积层的计算公式如（2-1）所示：

$$Y_{mn}^{(l+1)} = \sum_i^I \sum_j^J F_{ij}^{(l)} X_{(sm+i)(sn+j)}^{(l)} + b^{(l)} \quad (2-1)$$

其中, $F_{ij}^{(l)}$ 表示第 l 层网络中大小为 $I \times J$ 的卷积核, s 和 $b^{(l)}$ 分别表示卷积核移动的步长和第 l 层的偏置项, $Y_{mn}^{(l)}$ 表示第 l 层神经网络的输出特征图, $X_{mn}^{(l)}$ 表示第 l 层神经网络的输入图像或特征图。图 2-3 是一个简单的二维卷积示意图, 输入数据 I 经过一个尺寸为 3×3 、步长为 1 的卷积核 K 后输出数据 O。

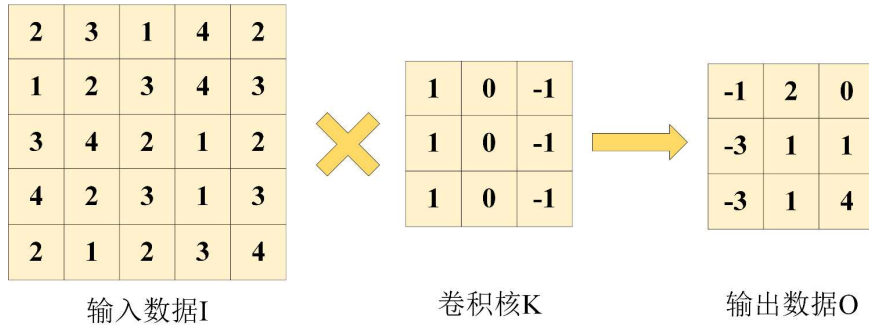


图 2-3 简单的二维卷积

2.3.2.2 池化层

在卷积神经网络的架构中, 过多的卷积层会大幅增加模型的参数量, 这不仅增加了计算的复杂性, 也提高了过拟合的风险^[60]。池化层的加入有效缓解了这一现象, 池化分为最大池化和平均池化^[61]。最大池化选取处理区域内最大的数值, 平均池化则取处理区域内所有值的平均值, 这两种池化都通过对特征图的子区域进行简化, 来生成更小尺寸的新特征图, 有效地降低了后续层的参数量和计算需求。两种池化的计算过程如图 2-4 所示。

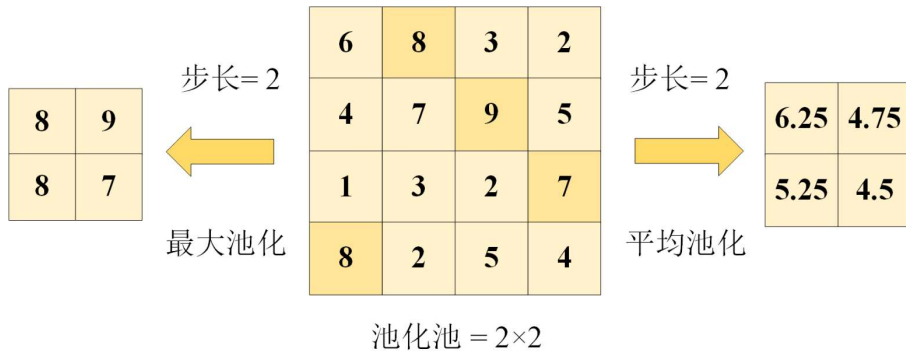


图 2-4 最大池化与平均池化计算过程

2.3.2.3 全连接层

全连接层通常作为网络模型的末端部分，位于卷积层和分类器之间，起到整合前面卷积层提取的特征并转换为用于最终任务的输出的作用。它通过将所有输入节点与输出节点相连接，实现特征的全面整合和映射，适用于分类、回归等多种任务。此外，全连接层也能实现特征降维，有助于简化模型结构，减少计算负担，同时保留有用信息，为决策提供有效支持。

2.3.2.4 激活函数

激活函数可以引入非线性特性，使得网络能够处理复杂的数据和模式，通常跟随卷积单元执行操作，处理卷积层生成的加权和。通过这种方式，激活函数能够转换卷积层的线性输出为非线性形式，增强网络的表达能力。这一过程对于网络学习和识别数据中的复杂模式和特征至关重要，有助于网络在进行图像分类、识别等任务时，更有效地捕捉和区分输入数据的细微差异。下面将对几种常用的激活函数进行介绍。

(1) Sigmoid 函数

Sigmoid 函数将输入映射到 (0, 1) 范围内，常用于二元分类。但在模型训练中易出现梯度消失问题，并且输出不是以 0 为中心，容易造成梯度消失和梯度爆炸。其计算方法如式 (2-2) 所示：

$$y = \frac{1}{1 + e^{-x}} \quad (2-2)$$

(2) ReLU 函数

ReLU (Rectified Linear Unit) 对正数输入返回相同的值，对负数输入返回 0。其计算简单，通常在深度学习中表现良好。但容易造成某些神经元输出永远为 0，导致网络部分失活。其计算方法如式 (2-3) 所示：

$$ReLU = \max(0, x) \quad (2-3)$$

(3) Leaky ReLU 函数

Leaky ReLU 是对 ReLU 的改进，其计算方法如式 (2-4) 所示：

$$LeakyReLU(x) = f(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha \cdot x, & \text{otherwise} \end{cases} \quad (2-4)$$

(4) Tanh 函数

Tanh 函数与 Sigmoid 函数类似，但其将输入映射到 (-1,1) 范围内，获得了

更大的输出范围，因此在某些情况下更适用，但仍然存在一定概率的梯度消失问题。其计算方法如式（2-5）所示：

$$\text{Tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (2-5)$$

（5）Softmax 函数

Softmax 函数通常用于多类别分类任务的输出层。但其生成的输出彼此之间相互影响，因此在多标签分类问题中可能不太适用。其计算方法如式（2-6）所示：

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \quad (2-6)$$

不同的损失函数具有不同的优缺点，选择适当的激活函数取决于目标任务的性质以及网络的结构，通常需要根据实际情况进行调整和选择。

2.3.2.5 损失函数

损失函数^[62]在深度学习模型的训练中扮演了核心角色，其主要功能是评估模型的预测值与实际标签之间的差异。这个差异度量指导着模型在训练过程中参数的更新，旨在通过最小化损失函数来优化模型的性能。损失函数不仅反映了模型完成特定任务的能力，比如分类或回归，而且能够考虑到数据的特性和任务的特性。此外，通过包含正则化项，损失函数还能帮助提升模型在未见数据上的泛化能力。下面将对几种常见的损失函数进行介绍：

（1）均方误差损失函数（Mean Squared Error, MSE）

均方误差损失是最常见的回归损失函数之一，其具体的计算方法如式（2-7）所示：

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (2-7)$$

（2）交叉熵损失函数（Cross-Entropy Loss）

交叉熵损失计算预测类别与真实类别之间的交叉熵，常用于神经网络的多分类任务。其具体的计算方法如式（2-8）所示：

$$\text{Cross-Entropy Loss} = -\frac{1}{N} \sum_{i=1}^N (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \quad (2-8)$$

（3）对数损失函数

对数损失通常用于二分类问题，在逻辑回归中广泛使用，计算真实标签和预测标签的对数损失。其计算方法如式（2-9）所示：

$$\text{Log Loss} = -\frac{1}{N} \sum_{i=1}^N (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \quad (2-9)$$

式（2-7）、（2-8）、（2-9）中的 y_i 是真实标签， $\hat{y}_i$ 是模型预测值， N 是样本数量。这些损失函数的计算方法可以根据具体的应用场景和问题类型进行相应的调整和扩展。

2.3.3 残差网络

随着卷积神经网络深度的增加，模型性能和计算精度通常会提升。然而，盲目增加网络深度会带来一系列问题。对于此，何等人^[63]提出了残差网络（ResNet）。ResNet 通过残差单元，其中包含直接映射和残差映射，有效避免了梯度消失和性能退化问题。残差单元的设计允许网络学习恒等映射，有助于深层网络的信息传递和训练，从而使得即使在网络极深时也能维持或提升模型性能，残差结构如图 2-5 所示。

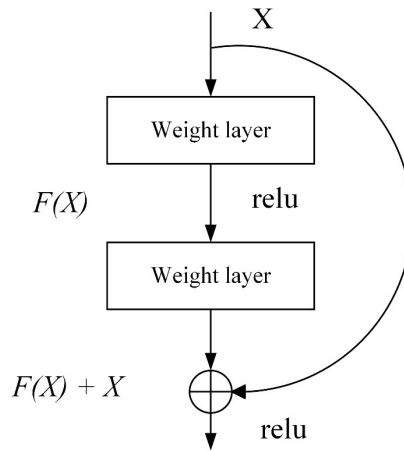


图 2-5 残差结构

在图 2-5 中， X 作为残差结构的输入，经过一系列卷积操作后，得到线性变化 $F(X)$ ， $F(X)$ 加上 X 后再经过激活函数输出，这里， $F(X)$ 就是要学习的残差。残差网络在普通神经网络的基础上引入了跳跃连接，有效的避免了由网络深度引起的梯度消失问题，使信息在多层网络间形成了传递，最大化的保留了特征信息。

2.3.4 注意力机制

注意力机制的设计来源于对人类视觉系统的理解。人类的眼睛无时无刻都在接收大量的信息，但并不是平等的对这些信息进行处理，大脑会在优先关注重要的信息。同理，在卷积神经网络中，尽管神经元在学习图像的特征时非常强大，但它们并不总能区分哪些特征比其它特征更重要。为了解决这个问题，研究者们引入了注意力机制，它允许网络模仿人类筛选重要特征的过程，从而提高了处理信息的效率和模型的性能。据此特性Hu等人^[64]提出了SENet，其基本结构如图2-6所示，该网络利用压缩与激励模块在通道维度上赋予特征图不同的权重，强化通道之间的映射关系，提升模型的整体性能^[65]。

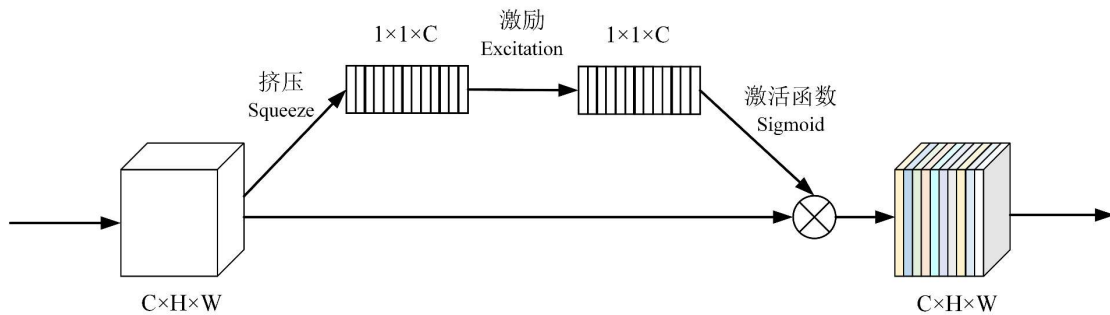


图 2-6 SENet 注意力机制

2.4 自编码器及其变体网络

2.4.1 自编码器介绍

自编码器（Autoencoder, AE）^[37]是一种有着多种用途的无监督网络模型，主要用于学习数据的高效表征或进行图像特征提取。如图2-7所示，核心架构由编码器和解码器两部分构成。具体而言，输入数据经过编码器处理后被转化为潜在空间中的低维表示，然后解码器利用这个低维表示重构原始输入数据，旨在学习数据的紧凑表示及其关键特征。

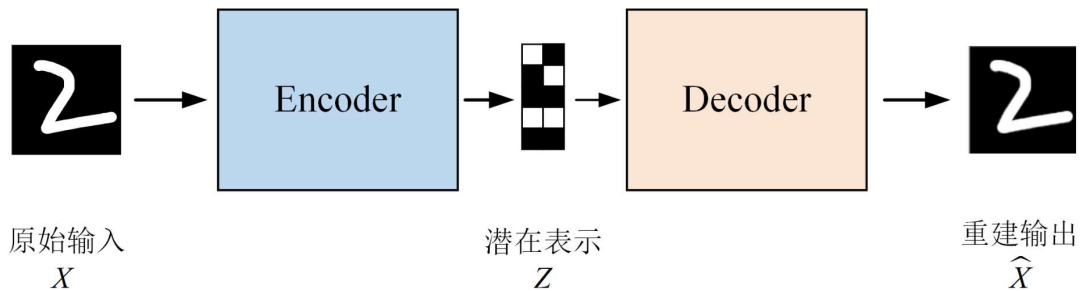


图 2-7 编码器的基本结构

自编码器具有多方面的应用，其中之一是在数据压缩领域。通过将数据压缩到潜在表示，可以对数据进行压缩和降维，减少存储空间和加速数据处理。此外，在特征学习领域，自编码器通过捕捉输入数据的关键特征，为监督学习任务提供了价值增强的输入，从而发挥着至关重要的角色。在去噪方面，自编码器通过学习重构无噪声的输出来从带噪声的数据中去除噪声，展示了其在数据清洗中的有效性。此外，作为一种生成模型，自编码器能够通过潜在空间内的采样及解码过程，生成具有相似属性的新数据样本。综上所述，自编码器是一种具有高度灵活性和广泛应用潜力的神经网络模型，适合解决多样的无监督学习任务。

自编码器因其独特的结构特性，在图像融合领域显示出显著的适应性。该模型通过学习输入图像的重要特征并将其转化为更低维的潜在表征，从而为图像融合过程提供了高效的前置输入。这一过程不仅实现了数据的降维与压缩，进而提升了图像融合的处理效率，而且其内置的去噪功能也有效减少了融合图像中可能出现的噪声干扰。作为一种生成型模型，自编码器能够基于融合的潜在表征生成全新的图像，为图像融合任务引入更多的创新性和多样性。得益于其无监督学习的特点，自编码器不依赖于大量的标注数据，使得它在实际应用场景中展现出极高的灵活性和实用价值。

2.4.2 自编码器网络结构

由图 2-7，自编码器的结构可以归纳为编码和解码两个部分，输入的数据通过编码器得到中间编码，解码器对其实施解码重构。如果将输入数据 X 映射为潜在表示 Z ，则可以将编码过程表示为：

$$Z = f_{enc}(X) \quad (2-10)$$

解码器将潜在表示 Z 映射回原始数据空间的重构数据 $\hat{X}$ ，则相应的可将解码过程表示为：

$$\hat{X} = f_{dec}(Z) \quad (2-11)$$

自编码器的目标是尽量使输入数据接近重构数据，通常使用式（2-7）提及的均方误差损失函数。而关于编码器和解码器的具体结构，根据输入数据格式的不同，可以采用神经网络中的多种算子，如全连接、一维卷积、二维卷积或三维卷积。同时，为了引入非线性能力，自编码器通常使用 Sigmoid、ReLU 和

Tanh 等激活函数。

2.4.3 其它变体网络

自编码器不仅能够学习数据的深层次表示，还能为其它机器学习任务提供丰富的特征信息，是深度学习领域的一个重要技术。相应的，根据不同任务，自编码器有许多不同的变体，包括基于全连接神经网络的自编码器、基于卷积神经网络的自编码器以及变分自编码器。

（1）基于全连接神经网络的自编码器

基于全连接神经网络的自编码器通过编解码器结构实现数据的有效表示学习和特征提取。其中，编码器通过利用自身多层的隐藏层结构对输入数据进行提取，进而将这些数据转换为潜在空间中的低维表示；解码器则对这些低维表示进行还原，通过特征提取和逐步升维恢复到原始数据空间。这个过程旨在学习数据的压缩表示，同时尽可能减少重构误差。

（2）基于卷积神经网络的自编码器

基于卷积神经网络的自编码器是专门为处理具有空间结构的数据设计的无监督学习模型。这种自编码器利用卷积层提取数据的空间特征，并通过池化层降低其空间维度，从而将输入数据压缩成一个更为紧凑的潜在表示。相应地，解码器部分通过反卷积层和上采样层将这一潜在表示逐步映射回原始数据的空间结构。这个过程不仅能够有效捕捉数据的重要特征，还能通过重构过程学习数据的深层结构和内容。基于卷积的自编码器因其在图像处理任务中的出色表现而广受欢迎，包括但不限于图像去噪、特征提取、图像增强和图像融合等应用场景，展现了其强大的功能和应用潜力。

（3）变分自编码器

变分自编码器（VAE）是一种先进的生成模型，它通过结合概率分布和潜在变量来揭示数据的内在结构。不同于传统自编码器，VAE 在映射过程中引入了条件概率分布，通过这种方式处理数据的不确定性，增强了模型的灵活性与可解释性。在训练 VAE 时，损失函数包含两个部分：重构损失评估重构数据与原数据之间的偏差。通过优化这一综合损失函数，VAE 不仅能够有效学习数据的深层结构，还能够生成新的、与训练数据相似的样本。VAE 在图像生成、文本生成和数据降维等多种领域展现了广泛的应用潜力，是无监督学习领域的一

种重要技术。为生成模型和无监督学习提供了一种强大的框架。

2.5 融合图像质量评价

图像融合领域缺少参考融合图像，通常采取主观肉眼判断和客观评价指标两种融合评估方法。但仅从一个方面评估融合结果并不科学，因此，在本文的融合实验中分别运用主观肉眼判断和客观评价指标对融合结果展开分析。

2.5.1 主观评价方式

主观评价方法依赖于人的直观感受来评判图像融合的效果，强调利用人眼对图像质量、色彩感知和信息保留度的准确判断。这种评价方式以图像的清晰度、信息保留、失真程度、对象完整性和视觉效果为标准，对不同融合技术产生的图像进行评价。其优点在于直观性和普遍性，能够在没有复杂计算的情况下提供快速反馈，有效节约计算资源。然而，主观评价受限于个人审美、视觉感知差异和显示设备质量，带有一定的不确定性和主观性，可能导致评价结果偏颇。当不同融合技术产生的图像差异微小时，仅依靠主观评价难以明确判断优劣，因此，在实际应用中应结合客观评价方法，确保融合效果的科学和可靠评估。

2.5.2 客观评价方式

相较于主观评价方法，客观评价提供了一种基于统计和计算的图像质量评估方式，避免了个人视觉偏好和外部条件的影响。客观评价通过一系列定量指标来衡量融合图像的质量，这些指标从不同维度反映了图像融合的效果和性能。但由于缺乏统一标准的参考融合图像和固定的评价体系，研究人员通常会选择多个指标来进行全面的融合效果评价。下面是 9 种常用的客观评价指标：

(1) 信息熵 (Entropy, EN) [66]

EN 可对融合图像中包含了多少来自源图像的信息量进行衡量，其计算过程如式 (2-12) 所示：

$$EN = - \sum_{i=0}^{L-1} p_i \cdot \log_2 p_i \quad (2-12)$$

其中 L 代表图像灰度级的数量， p_i 表示融合图像中灰度为 i 的像素所占融合图像总像素的比例。

(2) 平均梯度 (Average gradient, AG) [67]

AG 可对融合图像中纹理与细节的表征能力进行评测, 其计算过程如式 (2-13) 所示:

$$AG = \frac{1}{MN} \times \sum_{i=1}^M \sum_{j=1}^N \sqrt{\frac{1}{2} (f_x^2(x, y) + f_y^2(x, y))} \quad (2-13)$$

其中 f_x 和 f_y 分别代表图像在 x 和 y 方向上的梯度。

(3) 空间频率 (Spatial Frequency, SF) [68]

SF 可以反映从源图像到融合图像的灰度变化情况, 其计算方法如式 (2-14) 所示:

$$SF = \sqrt{\{RF\}^2 + \{CF\}^2} \quad (2-14)$$

其中 RF 和 CF 表示图像的空间行频率和列频率, 定义如 (2-15) 和 (2-16) 所示:

$$RF = \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (F(i, j) - F(i, j-1))^2} \quad (2-15)$$

$$CF = \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (F(i, j) - F(i-1, j))^2} \quad (2-16)$$

(4) 标准差 (Standard Deviation, SD) [69]

SD 可用于衡量融合图像中像素灰度值分布的离散程度, 它的计算方法如式 (2-17) 所示:

$$SD = \sqrt{\sum_{i=1}^M \sum_{j=1}^N (F(i, j) - \mu)^2} \quad (2-17)$$

其中 $F(i, j)$ 表示图像在点 (i, j) 处的像素灰度值, μ 为平均灰度值。

(5) 结构相似性 (Structural Similarity Index Measure, SSIM) [70]

SSIM 可从图像结构方面度量不同图像, 其计算方法如式 (2-18) 所示:

$$SSIM_{A,F} = \sum_{a,f} \frac{2\mu_a\mu_f + C_1}{\mu_a^2 + \mu_f^2 + C_1} \frac{2\sigma_a\sigma_f + C_2}{\sigma_a^2 + \sigma_f^2 + C_2} \frac{\sigma_{af} + C_3}{\sigma_a\sigma_f + C_3} \quad (2-18)$$

其中, $SSIM_{A,F}$ 表示图像 A 与图像 F 的相似性度量值, a 和 f 分别表示 A 和 F 的滑动区域大小, σ_{af} 表示的是 A 与 F 间的协方差, σ_a 和 σ_f 分别表示 A 与 F 的标准方差, μ_a 和 μ_f 分别表示为 A 与 F 的平均值。 C_1, C_2, C_3 均为参数。当存在源图像 A 和 B 时, 总的结构相似性定义如式 (2-19) 所示:

$$SSIM = SSIM_{A,F} + SSIM_{B,F} \quad (2-19)$$

(6) 互信息 (Mutual Information, MI) [71]

MI 是一种衡量两个变量之间相互依赖性的统计量, 其计算方法如式 (2-20) 所示:

$$MI = MI_{A,F} + MI_{B,F} \quad (2-20)$$

$MI_{A,F}$ 和 $MI_{B,F}$ 由式 (2-21) 计算得到:

$$MI_{X,F} = \sum_{x,f} p_{X,F}(x,f) \log_2 \frac{p_{X,F}(x,f)}{p_X(x)p_F(f)} \quad (2-21)$$

其中 $p_X(x) / p_F(f)$ 和 $p_{X,F}(x,f)$ 表示 X 和 F 的边缘概率分布函数和联合概率分布函数。

(7) 边缘信息度量 ($Q^{AB/F}$) [72]

$Q^{AB/F}$ 可用于测量融合图像对源图像边缘信息保留量, 它的计算方法如式 (2-22) 所示:

$$Q^{AB/F}(i,j) = \frac{\sum_{i=1}^M \sum_{j=1}^N (Q^{AF}(i,j)\omega^A(i,j) + Q^{BF}(i,j)\omega^B(i,j))}{\sum_{i=1}^M \sum_{j=1}^N (\omega^A(i,j) + \omega^B(i,j))} \quad (2-22)$$

其中 Q^{AF} 和 Q^{BF} 参考式 (2-23) 计算:

$$Q^{XF}(i,j) = Q_g^{XF}(i,j) \cdot Q_a^{XF}(i,j) \quad (2-23)$$

$\omega^A(i,j)$ 和 $\omega^B(i,j)$ 为权重参数, Q_g^{XF} 和 Q_a^{XF} 分别表示源图像 X 相对于融合图像 F 的相对强度和取向值, $M \times N$ 则表示图像的尺寸。

(8) 差异相关性和 (The sum of the correlations of differences, SCD) [73]

SCD 表示从源图像到融合图像的最大信息补充量, 其计算方式如下:

$$SCD = r(D_A) + r(D_B) \quad (2-24)$$

$$D_A(x,y) = F(x,y) - I_A(x,y) \quad (2-25)$$

$$D_B(x, y) = F(x, y) - I_B(x, y) \quad (2-26)$$

其中, D_A 与 D_B 分别代表两个源图像与融合图像之间的灰度值差异, $r(\cdot)$ 表示差异相关系数, $r(\cdot)$ 的具体计算过程如下:

$$r(D_A) = \frac{\sum_x^H \sum_y^W (D_A(x, y) - \bar{D}_A(x, y))(I_A(x, y) - \bar{I}_A(x, y))}{\sqrt{\left(\sum_x^H \sum_y^W (D_A(x, y) - \bar{D}_A(x, y))^2\right) \left(\sum_x^H \sum_y^W (I_A(x, y) - \bar{I}_A(x, y))^2\right)}} \quad (2-27)$$

$$r(D_B) = \frac{\sum_x^H \sum_y^W (D_B(x, y) - \bar{D}_B(x, y))(I_B(x, y) - \bar{I}_B(x, y))}{\sqrt{\left(\sum_x^H \sum_y^W (D_B(x, y) - \bar{D}_B(x, y))^2\right) \left(\sum_x^H \sum_y^W (I_B(x, y) - \bar{I}_B(x, y))^2\right)}} \quad (2-28)$$

(9) 视觉保真度指标 (Visual information fidelity of fusion, VIFF) [74]

VIFF 可用于评估融合图像对源图像的保真程度, 其计算如式 (2-29):

$$VIFF = \frac{\sum_{i \in \text{sub}} I(\overline{C^{N,l}}; \overline{F^{N,l}})}{\sum_{i \in \text{sub}} I(\overline{C^{N,l}}; \overline{E^{N,l}})} \quad (2-29)$$

其中, $\sum_{i \in \text{sub}} I(\overline{C^{N,l}}; \overline{F^{N,l}})$ 和 $\sum_{i \in \text{sub}} I(\overline{C^{N,l}}; \overline{E^{N,l}})$ 分别表示源图像与融合图像中相应自带分量的互信息。 $\overline{C^N}$ 表示随机场景中的 N 个元素信息, $\overline{F^N}$ 、 $\overline{E^N}$ 则表示源图像与融合图像输出的 HSV 模型。

上述 9 种评价指标为评估融合图像的质量提供了一种相对客观的方法, 并且这些指标的数值通常与融合质量的好坏正相关。也就是说, 指标值越大, 融合质量越好。

2.6 本章小结

本章主要对图像融合领域涉及的相关基础知识进行了介绍。首先, 对红外与可见光图像的成像原理及特点进行了介绍; 其次, 介绍了图像融合的几个层级; 然后, 对深度学习和自编码器网络的相关知识进行了概括总结; 最后, 总结了评价图像融合成像质量的相关依据, 包括主观视觉观察方法以及客观评价指标方法, 为后续的实验融合结果质量比较提供了充足的理论基础。

第3章 基于双尺度编码结构与注意力机制的融合算法

3.1 引言

本章针对现有融合算法常从单一尺度提取特征，导致融合图像中红外图像特征与可见光图像特征表现不全面、不均衡等问题，提出一种基于双尺度编码结构与注意力机制的融合算法 DSAFusion。算法首先在使用八度卷积替换普通标准卷积对源图像进行双尺度划分的基础上，采用密集链接模块和语义信息模块构建特征提取子网络；接着，在编码网络的末端引入注意力机制 CoTNet 捕捉全局信息的依赖关系，强化重点特征表达；然后，把通道注意力机制与空间注意力机制引入融合策略，凸显重要特征；最后，融合后的特征被传递进解码网络中并重组成信息丰富的融合图像。将 DSAFusion 分别在 TNO 和 LLVIP 数据集上进行融合测试实验，并与现有 7 种优秀的融合算法进行了比较。在主观视觉评价方面，DSAFusion 生成的融合图像在保留了较多细节纹理信息的同时加强了显著目标信息的表达，保证了红外图像特征与可见光图像特征之间的平衡，整体观感清晰，符合视觉感知；在客观评价指标方面，选择 6 个评价指标，该算法在 TNO 数据集上的熵 (En)、空间频率 (SF)、结构相似性 (SIMM) 和视觉信息保真度 (VIF) 4 个指标上取得了最优，在 LLVIP 数据集上同样取得了这 4 个指标的最优表现。实验结果表明，DSAFusion 具有优异的图像融合能力与融合效率。

3.2 融合网络介绍

3.2.1 八度卷积

八度卷积 (Octave Convolution) ^[75] 与传统的标准卷积不同，它设计了一种新的方法来同时处理高频和低频信息，通过添加一个额外的分支来专门处理低频信息，可在处理图像时更细致地区分不同频率的特征。在八度卷积中，高频特征图的空间维度是低频特征图的两倍，这样的设计允许模型在减少信息损失的同时有效降低数据的空间复杂度，从而实现更高效的特征提取和信息处理。

其详细的结构如图 3-1，输入特征图的初始高度和宽度分别由 H 和 W 表示，而输入和输出的通道数分别由 C_{in} 和 C_{out} 表示。第一行中的 X^H ，即八度卷积

模块的原始输入特征图；第二行的 X^L ，表示经过缩减尺寸处理的低频特征图。这里， α_{in} 作为比例系数，定义了输入低频特征图通道数占总输入通道数的比例，其值在 0 到 1 之间变化。 α_{out} 与 α_{in} 有着相同的定义，用于表示输出低频特征图通道数占总输出通道数的比例，其取值范围也是 0 到 1。特别地，当 α_{in} 和 α_{out} 均设为 0 时，八度卷积将转变为常规的标准卷积。网络中还包括“Conv”（标准卷积）、“Pooling”（自适应均值池化）和“Upsampling”（特征上采样）等关键操作，这些都是实现高效图像处理的基础。

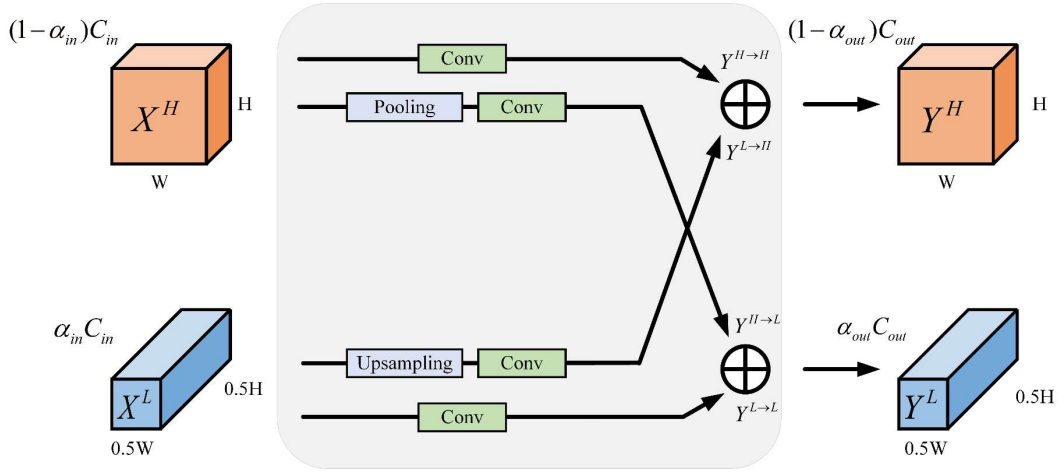


图 3-1 八度卷积

3.2.2 注意力机制 CoTNet

传统的自注意力机制主要通过独立的查询机制和“key-value”对来学习注意力矩阵，但这种方法往往忽略了变量间的丰富上下文信息。DSAFusion 引入的自注意力模块 CoTNet^[76]结构如图 3-2 所示。其中，输入的特征图表示为 X ，高、宽和通道数分别表示为 H 、 W 、 C 。CoTNet 首先通过一个 3×3 尺寸的卷积层处理 X ，生成特征 K_1 ，该特征被视为静态特征。同时， X 也通过 1×1 尺寸的卷积层生成向量 $V(x)$ 。为了增强特征间的互动，将 K_1 与原始特征 X 级联，特征通道数增加至 $2C$ 。接着，通过两个 1×1 卷积层的连续处理，并与 $V(x)$ 相乘，产生全局动态特征变量 K_2 。最终，通过融合局部静态向量 K_1 与全局动态变量 K_2 ，得到增强特征 Y 。CoTNet 通过将上下文信息提取和注意力学习融为一体的架构，有效的对红外与可见光图像中的显著性特征进行了保留。

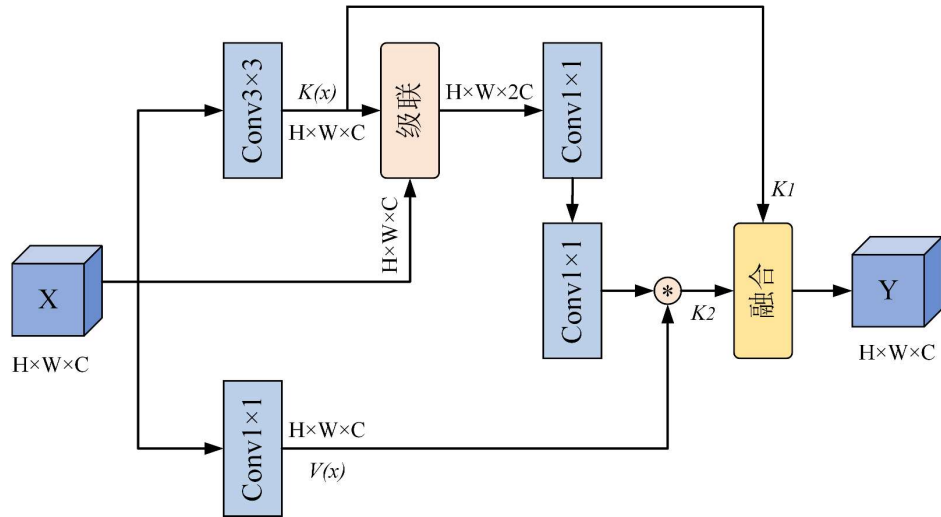


图 3-2 注意力机制 CoTNet

3.2.3 主体网络结构

DSAFusion 首先由双尺度编码器提取红外与可见光图像的特征，然后经融合层执行融合操作，最后在解码器部分进行图像重建。DSAFusion 的整体网络结构如图 3-3 所示。

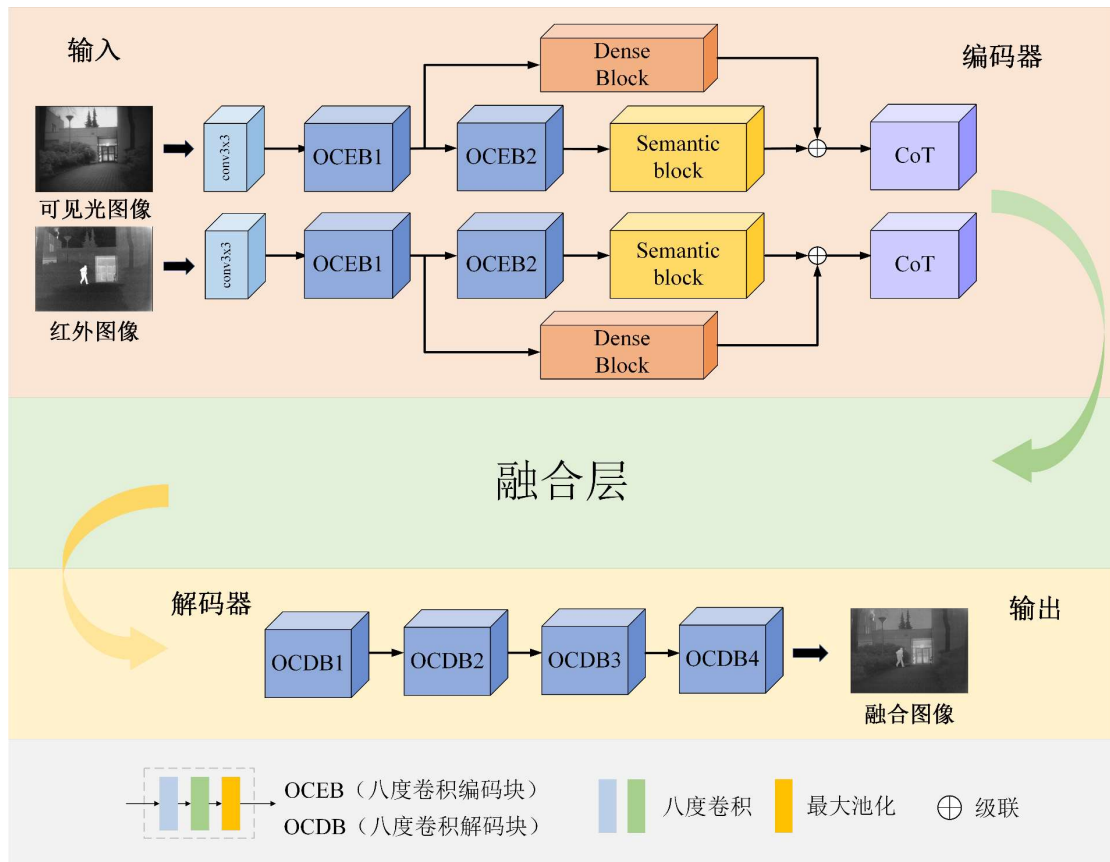


图 3-3 DSAFusion 算法结构

其中，网络的输入为若干尺寸调整至 256×256 的红外与可见光图像对，它们首先经由一个卷积核为 3×3 的标准卷积被输入进两层八度卷积网络中。通过第一层八度卷积（OCEB1）后获得了 16 个通道特征，将这些通道特征视为浅层特征，并在第一个八度卷积后接入密集链接模块（Dense Block），密集链接模块的具体结构如图 3-4 所示。由于浅层卷积提取的特征往往包含更详细的细节纹理信息。因此，DSAFusion 中的密集链接模块包含三层卷积，每一层卷积的输入均与所有先前层的输出级联。在这些卷积中，特征的尺寸不发生改变，而通道数累加，分别为 32、48、64。这种密集链接网络持续的强化了源图像底层的细节信息，使编码结构能够更好的提取浅层特征，有效地对源图像的背景纹理信息进行了保留。

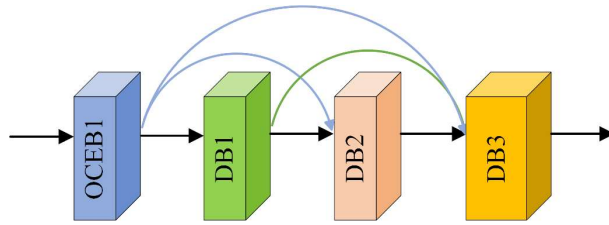


图 3-4 密集链接网络结构

接着经过第二层八度卷积（OCEB2）将获得 32 个通道特征，将这些通道特征视为深层特征，并在其后接入语义信息模块（Semantic Block），语义信息模块的具体结构如图 3-5 所示。相对于浅层卷积，深层卷积拥有更大的感受野，提取的特征往往传递了更多的语义信息。因此，DSAFusion 的语义信息模块先对深层特征进行下采样，以步长为 2 执行三次卷积运算使深层特征的尺寸减小到 128×128 、 64×64 和 32×32 。并同步将深层特征的通道数增加至 64 和 128，再减小至 64。使语义信息模块最终的通道数与密集链接模块最终的通道数相同，以便经过语义信息模块的深层特征还能与经过密集链接模块的浅层特征能相连接。最后，在语义信息模块的末端，利用双线性插值对深层特征进行上采样操作，保持通道数不变，将它们的大小扩展至 256×256 。经过这样快速的尺寸调整，图像特征中较为不明显的背景细节信息将会大量丢失，剩下较为明显的显著目标信息，极大程度的避免了特征冗余，防止融合图像出现伪影。

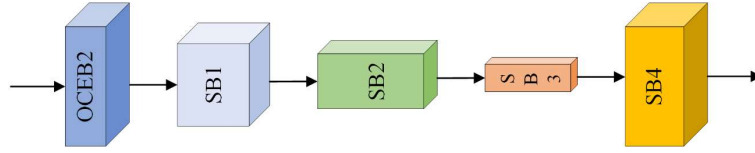


图 3-5 语义信息网络结构

随后，先将从密集链接模块获取的浅层特征与从语义信息模块获取的深层特征在同一特征空间中进行级联。接着，这些合并的特征被送入 CoTNet 注意力机制模块，增强图像的视觉表现力。经过 CoTNet 处理的特征随后被输入到融合网络，以产生融合后的特征。这些融合特征最终被送入解码器中执行图像重建步骤。解码器由四个八度卷积块组成，其输出通道数依次为 64、32、16 和 1。通过最后一个 OCDB4 卷积块重建出最终的融合图像。此外，为提升网络的计算效率，这些卷积块均未采用批量归一化层。

3.2.4 融合策略

DSAFusion 通过在特征融合阶段融入空间和通道注意力机制，实现了对特征重要性的自适应学习。具体而言， Φ_{ir} 和 Φ_{vi} 分别代表从编码网络中提取的红外和可见光图像特征。由于这两种特征都是通过相同的网络和参数提取的，它们可以直接相加以获得初步的融合特征 Φ_F^1 。接下来， Φ_F^1 经过通道注意力机制的作用，首先经过全局平均池化（GAP）转换成特征向量，然后利用两个全连接层（FC）学习不同特征通道之间的相互依赖性，接着通过 Sigmoid 激活函数计算得到的权重值介于 0 和 1 之间，这些权重值乘以 Φ_F^1 得到 Φ_F^2 。最后， Φ_F^2 通过空间注意力机制进行进一步处理，使用 1×1 卷积核的标准卷积生成投影张量，并通过 Sigmoid 函数进行空间维度上的调整。调整后的特征与 Φ_F^2 相乘，产生最终融合特征 Φ_F^3 。

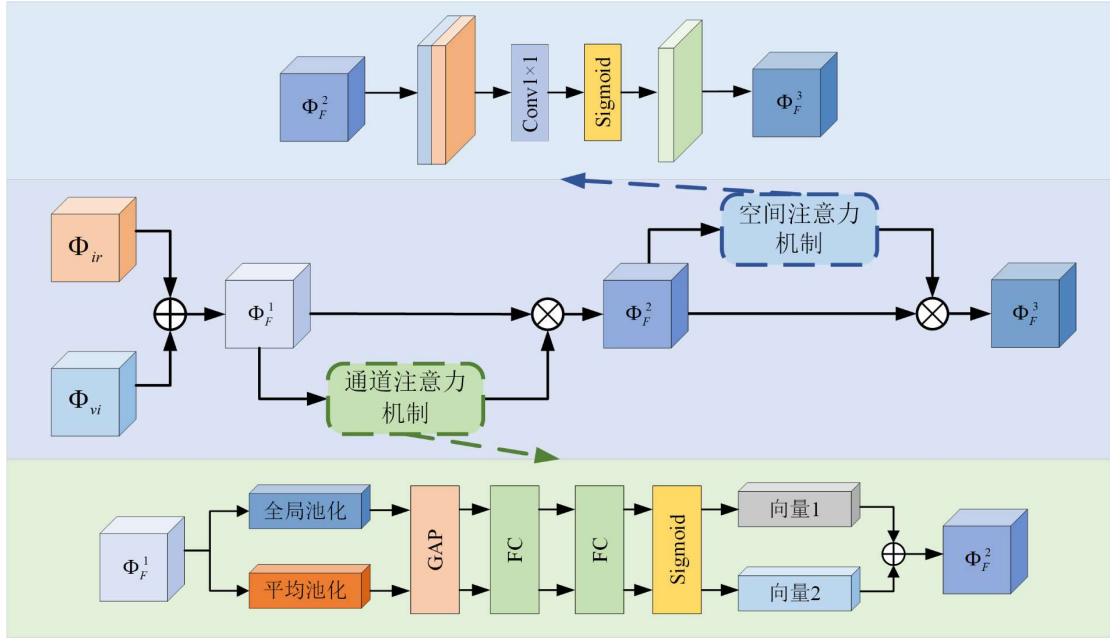


图 3-6 DSAFusion 融合网络结构

3.2.5 训练阶段

DSAFusion 需要训练一个自编码网络（AE）^[37]，为了优化训练过程，训练时从整体网络结构中移除了融合层，这种做法允许编码器从多个尺度提取图像特征，而这些特征在经过注意力机制加工后，由解码器进行重建工作。这个流程确保了网络能够学习到如何有效地集成来自不同层次的信息，从而提高最终融合图像的质量和相关性。训练阶段的网络结构如图 3-7 所示。

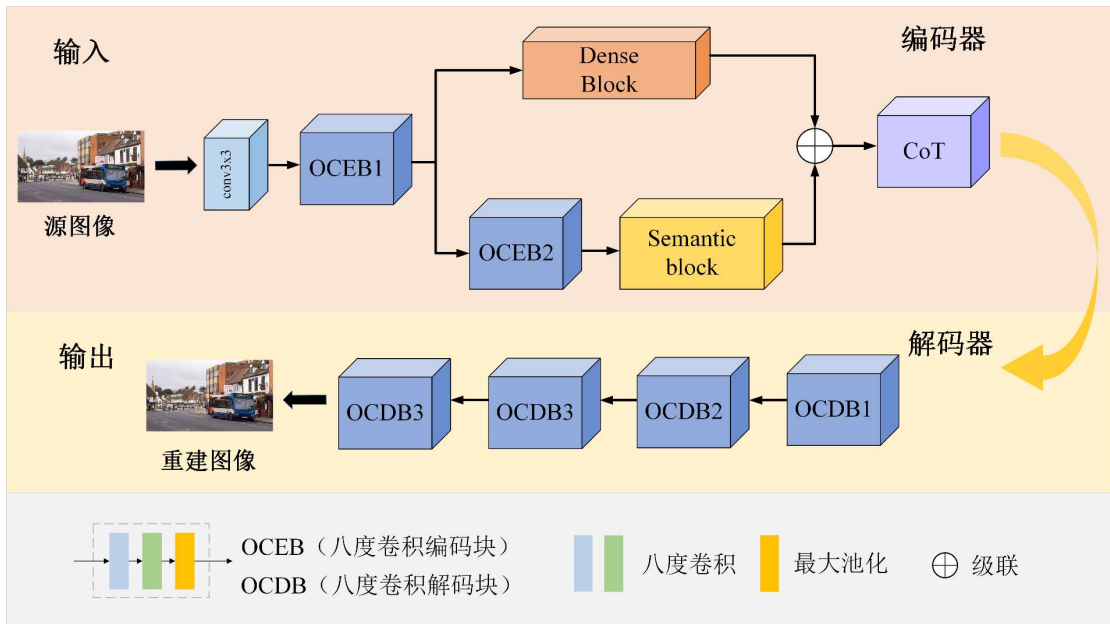


图 3-7 训练阶段网络结构

图 3-7 的输入是一幅用于训练的源图像，而输出是由解码器还原重建出的图像。DSAFusion 使用损失函数 L_{AE} 进行训练，采用像素损失和结构相似性损失^[70]作为重构损失，具体计算为：

$$L_{AE} = L_{Pixel} + \lambda L_{SSIM} \quad (3-1)$$

式 (3-1) 中， L_{AE} 、 L_{Pixel} 和 L_{SSIM} 分别代表图像重建损失、像素损失和结构相似性损失， λ 则是用来平衡像素损失和结构相似度损失的超参数。 L_{Pixel} 的计算公式为：

$$L_{Pixel} = \|I_{re} - I_{in}\|_F^2 \quad (3-2)$$

式 (3-2) 中， I_{re} 和 I_{in} 分别代表重构图像和源图像， $\|\cdot\|_F^2$ 表示 Frobenius 范数。像素损失可以使重构图像在像素级别上更接近源图像。此外，结构相似度损失计算公式为：

$$L_{SSIM} = 1 - f_{SSIM}(I_{in}, I_{re}) \quad (3-3)$$

在训练阶段，DSAFusion 仅使用编码器提取图像特征，并不需要成对的红外与可见光图像进行训练。本小节利用 MS COCO 数据集^[77]对自编码网络进行训练。又因为红外图像为单通道的灰度图像，所以在训练时还需要将数据集中的彩色图片进行灰度处理。这一过程通过对每个像素点的红、绿、蓝 (RGB) 值进行加权平均来实现，其中常用的加权公式如 (3-4) 所示：

$$Gary = 0.299R + 0.587G + 0.114B \quad (3-4)$$

通过色阶转换，彩色图像中每个像素点的 RGB 值被转换为一个单一的灰度值，从而形成与红外图像相同的灰阶图像，图 3-8 为转换结果的示例图像。

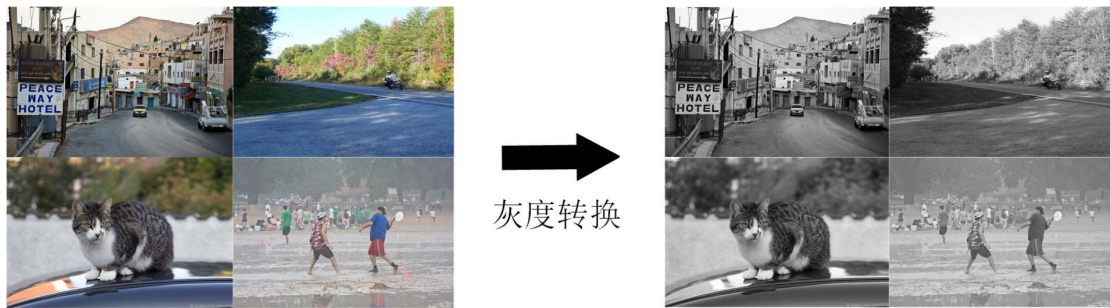


图 3-8 灰度转换示例

转换为灰度图像后，将这些图片的尺寸调整至 256×256 ，使用 Adam 对网络进行优化，并将 lr 的大小设置为 1×10^{-4} ，Batchsize 的大小设置为 2，进行 8 个 Epoch 的训练。网络的一些参数设置见表 3-1。

表 3-1 DSAFusion 具体网络参数

		卷积层	卷积核	步长	输入通道数	输出通道数	激活函数
编码器 (Encoder)	卷积层	Conv3×3	3×3	1	1	8	ReLU
	编码器	OCEB1	3×3	1	8	16	ReLU
	八度卷积块	OCEB2	3×3	1	16	32	ReLU
	密集链接模块	DB1	3×3	1	16	32	ReLU
	(Dense	DB2	3×3	1	32	48	ReLU
	Block)	DB3	3×3	1	48	64	ReLU
	语义信息模块	SB1	3×3	2	32	64	ReLU
	(Semantic	SB2	3×3	2	64	128	ReLU
	Block)	SB3	3×3	2	128	64	ReLU
		SB4	3×3	2	64	64	ReLU
解码器 (Decoder)		OCDB1	3×3	1	128	64	ReLU
	解码器	OCDB2	3×3	1	64	32	ReLU
	八度卷积块	OCDB3	3×3	1	32	16	ReLU
		OCDB4	3×3	1	16	1	ReLU

3.3 融合实验及效果评估

3.3.1 实验环境及评价指标

DSAFusion 算法的所有实验均在操作系统 Ubuntu18.04 上进行，CPU 为 Intel i7-11800H，GPU 为 NVIDIA RTX 3060，编程语言为 Python3.6，深度学习的框架为 Pytorch。训练阶段与 3.2.5 小节中介绍一致，测试集选用 TNO^[48]和 LLVIP^[49]。由于可见光图像也为三通道的彩色图像，所以在正式实验前，还需要对彩色的可见光图像进行灰度处理。处理后的图像样本如图 3-9 所示：

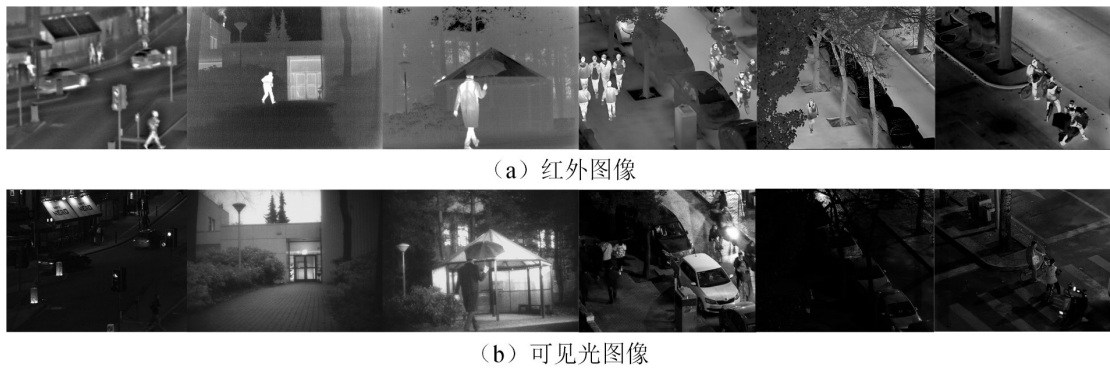


图 3-9 灰度转换后的数据集图例

对所提算法的融合结果从主观视觉和客观指标两个方面进行评价, 选择 7 种现有的优秀图像融合算法进行比较, 包括 NSCT^[25], LatLRR^[78], CNN^[29], FusionGAN^[33], DDcGAN^[34], DenseFuse^[38]以及 NestFuse^[39]。这些算法的代码及实验配置均从原参考论文中获取。客观方面则从 2.4.1 小节中选择了 6 个评价指标来进行衡量, 包括 SCD^[73], AG^[67], SF^[68], EN^[66], SSIM^[70]和 VIFF^[74]。融合图像的质量均随着这六个评价指标数值的增加而提高。

3.3.2 参数选择

在 3.2 小节中详细介绍了所提方法中使用的八度卷积模块, 为了确定八度卷积块中合适的 α_{in} 、 α_{out} 取值, 先将它们设置为 0.125、0.25 和 0.5, 进行相关实验。此外, 按照 3.2.5 小节, 同步将损失函数里的平衡参数 λ 的值分别设置为 10、100 和 1 000 进行参数交叉实验。

在当前实验中, 编码器、融合网络以及解码器的结构均与前文中介绍的结构保持一致。利用 3.3.1 小节中提及的 6 个客观评价指标来评估具有不同 α_{in} 、 α_{out} 以及不同 λ 取值的融合网络性能。融合结果的客观指标如表 3-2 所示, 其中粗体表示最优值, 斜体表示次优值, 后续表格均用此标注表示。

表 3-2 不同参数网络融合结果评价指标平均值

α_{in}/α_{out}	λ	EN ^[66]	AG ^[67]	SF ^[68]	SCD ^[73]	SSIM ^[70]	VIFF ^[74]
0.125	10	3.89	2.997	6.98	29.46	0.533	0.479
0.125	100	6.91	4.669	9.56	35.34	0.759	0.681
0.125	1000	6.28	4.199	8.74	37.25	0.671	0.613
0.25	10	3.37	3.482	5.92	29.17	0.524	0.377
0.25	100	5.33	4.468	8.22	33.91	0.687	0.598
0.25	1000	4.82	3.876	7.23	32.86	0.598	0.501
0.5	10	2.73	2.066	4.27	27.33	0.289	0.218
0.5	100	3.92	3.609	5.72	29.89	0.429	0.472
0.5	1000	5.75	4.179	7.59	33.11	0.516	0.523

由表 3-2 可以看出, 当 α_{in} 和 α_{out} 设置为 0.125, λ 设置为 100 时, 网络获得了最多的最优值评价指标, 即网络拥有更优的性能。所以在接下来的实验中, 除 OCEB1 的 α_{in} 与 OCDB4 的 α_{out} 设置为 0, 其余所有 α_{in} 和 α_{out} 的取值均设置成 0.125, 并将训练阶段损失函数的平衡参数设置为 100。

3.3.3 实验结果分析

将所提算法与 3.3.1 节提及的现有的优秀的融合方法进行对比,不同方法的融合结果对比如图 3-10、图 3-11 以及图 3-12、图 3-13 所示,其中图 3-10、图 3-11 实验结果取自 TNO 测试集^[48],图 3-12、图 3-13 实验结果取自 LLVIP^[49]数据集。分别用红色实线框和黄色虚线框标记出红外图像中的显著目标信息与可见光图像中的细节纹理信息。并用相应颜色的矩形框给出局部放大图。

从图 3-10、图 3-11 融合结果中可以看出: NSCT^[25]算法、LatLRR^[78]算法和 FusionGAN^[33]算法的融合结果与红外图像的视觉效果极为相似,这表明这些算法在融合过程中未能有效捕获来自可见光图像的背景纹理信息。特别是在图 3-11 中,车辆的边缘模糊不清,说明这些算法在保留细节方面存在不足。另一方面, DenseFuse^[38]算法和 NestFuse^[39]算法表现出了对可见光图像中细节信息的过度强调,以至于抑制了红外图像中重要的显著目标信息的展现。在图 3-11 中,实线框中的人物轮廓十分模糊,反映了这些算法在平衡红外图像和可见光图像信息方面的局限。与此同时, CNN^[29]算法、DDcGAN^[34]算法和 DSAFusion 算法在融合图像时表现出了很好的平衡。在图 3-10 和图 3-11 中,红色实线框内的显著目标信息和黄色虚线框内的细节信息均被清晰地保留。然而, CNN 算法生成的融合图像表面出现了大量噪声,破坏了图像的整体观感。DDcGAN^[34]算法虽然避免了噪声问题,但其生成的图像清晰度不足,导致整体图像看起来模糊不清。在这些算法中, DSAFusion 算法展现了独特的优势。 DSAFusion 在进行源图像特征提取时,能够根据不同的尺度动态地利用相应的网络模块。该方法生成的融合图像不仅均衡地保留了红外图像中的显著目标信息和可见光图像中的背景纹理信息,还避免产生了多余的噪声,确保了融合图像的清晰度。通过这种细致的特征处理和信息平衡, DSAFusion 算法能够产生既清晰又内容丰富的融合图像,优于其它算法在图像质量和信息表达上的表现。

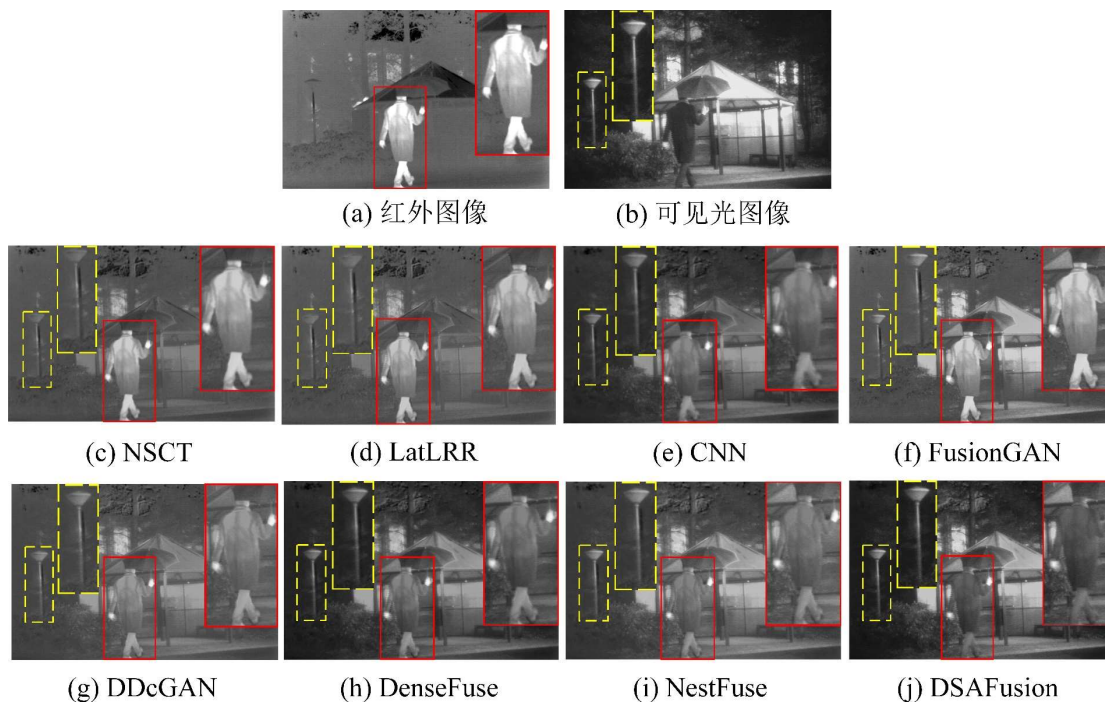


图 3-10 不同算法在 TNO 测试集上的融合结果一

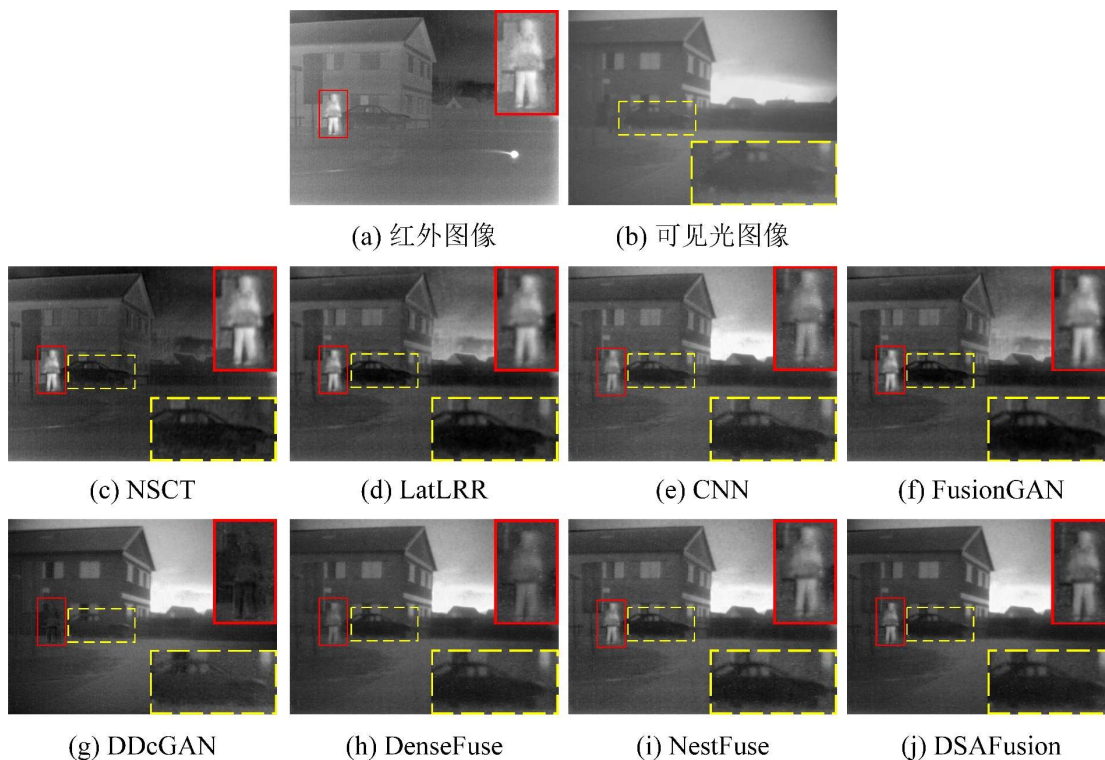


图 3-11 不同算法在 TNO 测试集上的融合结果二

LLVIP 测试集与 TNO 测试集相比, 提供了更加丰富和复杂的户外监控场景, 其中目标众多且背景杂乱, 对融合效果的验证十分有意义, 测试结果如图 3-12、图 3-13 所示, 在这些具有挑战性的场景下, 准确的特征提取对于融合效果至关

重要。同 TNO 测试集表现一致，NSCT^[25]算法、LatLRR^[78]算法和 FusionGAN^[33]算法成功突出了红外图像中的显著目标信息，但在保留可见光图像的背景纹理方面表现不佳。例如，在图 3-12 中，车辆内部的细节信息几乎无法辨认，反映了这些算法虽然能够强调重要目标，但在处理复杂背景时缺乏精度。相反，DenseFuse^[38]算法和 NestFuse^[39]算法在尝试保留可见光图像中的细节时，却过度抑制了红外图像的显著目标信息，导致如图 3-12 中人群和图 3-13 中行人的特征变得模糊不清。凸显了这些算法在平衡显著目标与背景细节方面的挑战。CNN^[29]算法和 DDcGAN^[34]算法生成的融合图像在某种程度上出现了泛白现象，影响了图像的整体观感。特别是 DDcGAN^[34]算法在维持背景细节时表现不尽人意，导致图像的信息损失。另一方面，DSAFusion 算法引入的注意力机制在复杂的户外监控场景下展现出其优势，通过有效抑制不重要的信息而保留关键特征，不仅提升了融合图像的质量，还优化了视觉观感，更贴近人类视觉感知系统。同时保留了重要目标信息，产生了清晰信息丰富的融合图像，表现出其优异的融合能力。

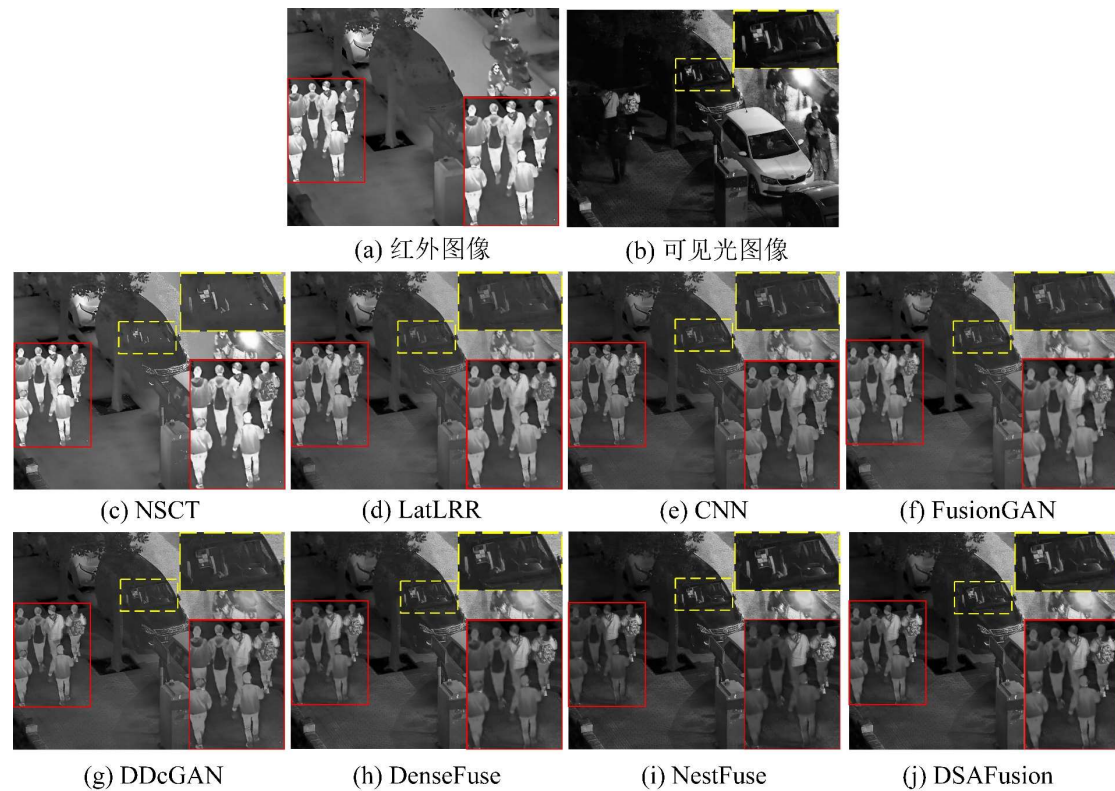


图 3-12 不同算法在 LLVIP 测试集上的融合结果一

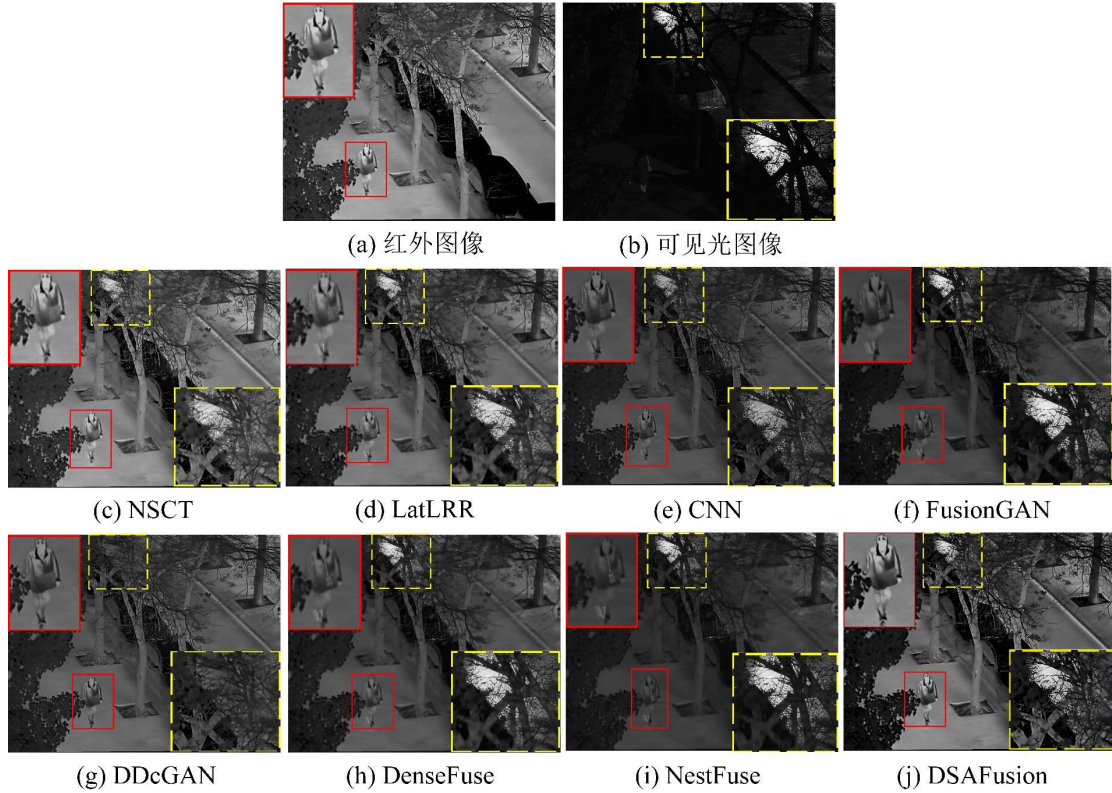


图 3-13 不同算法在 LLVIP 测试集上的融合结果二

综合以上实验结果分析，可以明显看到，除 **DSAFusion** 外的其它图像融合方法在特征表达的均衡性、融合图片的模糊度以及清晰度方面均存在一定程度的不足。相对而言，**DSAFusion** 算法在融合红外图像的显著目标信息与可见光图像中的细节纹理方面表现出了显著优势。这种优势主要归功于其特征提取网络中有效的尺度划分以及注意力机制的引入。合理的尺度划分允许 **DSAFusion** 算法在处理不同尺度的特征时，能够更加精细地捕捉到图像中的重要信息。同时，注意力机制的作用确保了它能够专注于那些有益于融合效果的关键特征，从而更好的展现在最终的融合图像中。

主观评价虽然能够提供直观的感受，但因其与个人的审美标准和偏好紧密相关，可能会导致评价结果的主观性和不一致性。特别是当多种融合方法的效果相近时，仅凭肉眼判断优劣变得尤为困难。因此，采用 3.3.1 节中选择的 6 个客观评价指标对这些融合方法进行比较。在 TNO 测试集上的实验结果如表 3-3 所示，LLVIP 测试集上的结果如表 3-4 所示。这种基于客观评价指标的比较方法，不仅提供了一种量化的、标准化的评估手段，也使得不同融合方法之间的比较更加明确和具有可比性。

表 3-3 不同算法在 TNO 测试集试验结果评价指标平均值

融合算法	EN ^[66]	AG ^[67]	SF ^[68]	SCD ^[73]	SSIM ^[70]	VIFF ^[74]
NSCT ^[25]	5.98	2.213	6.76	37.78	0.633	0.356
LatLRR ^[78]	6.43	2.737	7.34	40.95	0.687	0.384
CNN ^[29]	6.24	4.175	8.98	30.6	0.735	0.493
FusionGAN ^[33]	6.36	2.954	6.91	33.06	0.683	0.456
DDcGAN ^[34]	5.86	3.678	7.79	29.88	0.716	0.533
DenseFuse ^[38]	6.17	3.803	7.44	23.54	0.727	0.596
NestFuse ^[39]	6.84	5.381	8.36	31.27	0.692	0.550
DSAFusion	6.91	4.669	9.56	35.34	0.759	0.681

表 3-4 不同算法在 LLVIP 测试集试验结果评价指标的平均值

融合算法	EN ^[66]	AG ^[67]	SF ^[68]	SCD ^[73]	SSIM ^[70]	VIFF ^[74]
NSCT ^[25]	5.74	2.153	10.87	38.34	0.702	0.512
LatLRR ^[78]	5.92	3.085	11.24	36.56	0.769	0.554
CNN ^[29]	6.25	4.568	13.98	27.66	0.834	0.696
FusionGAN ^[33]	5.60	2.594	12.97	29.71	0.748	0.334
DDcGAN ^[34]	5.34	3.249	12.72	31.22	0.729	0.717
DenseFuse ^[38]	6.25	4.163	12.87	24.04	0.826	0.669
NestFuse ^[39]	6.52	5.669	13.56	25.07	0.857	0.637
DSAFusion	6.73	5.478	14.29	37.28	0.883	0.745

由表 3-3 可以看出, 在 TNO 测试集上, DSAFusion 与其它融合算法相比取得了 4 个指标上的最优。其中, 熵 (EN^[66]) 和空间频率 (SF^[68]) 的最优代表图像特征在空间中更为活跃, 说明了所提方法可从源图像中提取出更多的有用信息; 结构相似性 (SSIM^[70]) 和视觉信息保真度指标 (VIFF^[74]) 则代表融合图像更接近源图像, 证实了 DSAFusion 的融合结果更加清晰, 观感更好。而表 3-4 的实验结果显示, DSAFusion 在不同的数据集上表现基本一致。在 LLVIP 数据集上, 同样取得了熵 (EN^[66])、空间频率 (SF^[68])、结构相似性 (SSIM^[70]) 和视觉信息保真度指标 (VIFF^[74]) 四个指标的最优。这也说明了本章 DSAFusion 算法模型在面对不同的数据集时具有良好的泛化能力。

综上所述, 所提 DSAFusion 算法得益于在两个不同的尺度上对应连接了不同的网络模块以及注意力机制的作用, 在对源图像细节纹理信息的保留与显著性增强方面存在优势。对于融合结果, 无论从主观肉眼判断, 还是客观的评价指标分析, 本章算法都可在保留源图像提供丰富信息的基础上, 有更好的清晰

度和更强的对比度，实现红外和可见光图像信息的良好平衡。

3.3.4 消融实验

为了验证 DSAFusion 算法中双尺度模块和注意力机制 CoTNet^[76]的作用，本节利用 TNO 数据集设计了消融实验，采用客观评价法进行实验结果测评。其中，针对双尺度模块，分别设置“仅 Dense Block 分支”和“仅 Semantic Block 分支”两种网络变体进行试验。而对 CoTNet 则直接观察它的加入对网络的影响。这里，除编码器中的网络模块和注意力机制 CoTNet 外，其余如学习率、训练轮数、比例系数、损失函数平衡参数等配置均固定且与前文一致，实验结果如表 3-5。

表 3-5 双尺度模块与注意力机制的消融实验

实 验	Dense Block	Semantic Block	CoTNet ^[76]	EN ^[66]	AG ^[67]	SF ^[68]	SCD ^[73]	SSIM ^[70]	VIFF ^[74]
1	√	-	√	4.83	4.324	8.23	36.15	0.519	0.434
2	-	√	√	5.35	3.689	8.93	29.98	0.536	0.419
3	√	√	-	4.56	3.980	7.37	30.76	0.687	0.585
4	√	√	√	6.91	4.669	9.56	35.34	0.759	0.681

由表 3-5 数据可知，当三者共同作用时，实验数据要明显优于其它情况，也证明了双尺度模块和注意力机制 CoTNet 共同作用的有效性。此外，为了验证八度卷积是否能提高网络的运行效率，在不同数据集、不同图像尺寸的情况下，将本章 DSAFusion 算法的平均运行时间与融合结果相似的 DDcGAN^[34]、DenseFuse^[38]和 NestFuse^[39]相比较，结果如表 3-6 所示。

表 3-6 不同算法对 40 个图像对的平均处理时间

融合方法	DDcGAN ^[34]	DenseFuse ^[38]	NestFuse ^[39]	DSAFusion	DSAFusion*
平均运行时间(S)	1.415	0.235	0.318	0.162	0.213

表 3-6 中 DSAFusion*指将八度卷积替换回标准卷积的 DSAFusion，由于网络结构较为简单，本章算法有着最快的运行时间，而八度卷积的加入更是进一步降低了内存的消耗，加快了融合网络的响应速度。

3.4 本章小结

本章针对常规融合网络常从单一尺度提取特征的局限，提出了一种基于双尺度模块与注意力机制的融合算法 DSAFusion。首先利用由八度卷积组成的网

络对源图像进行了双层的特征提取，在第一层后连接了密集链接模块以保留更多的细节纹理信息，在第二层后连接了语义信息模块以强化显著目标信息；接着把两组特征在一个特征空间内级联后输入进注意力机制 CoTNet 中，感知静态向量与动态上下文信息，抑制无用信息表达；然后将其输入进融合网络中进行特征融合；最后由解码器将融合后的特征重构成一幅融合图像。DSAFusion 在 GPU 上基于 PyTorch 框架实现，并在不同的数据集上与不同现有融合方法进行了比较，均取得了较为优秀的融合结果，说明该算法可以充分有效的提取源图像中的特征信息，并且保证了融合图像有着较高的清晰度，更加符合人眼的视觉感知。

第4章 基于多尺度编码结构与特征增强模块的融合算法

4.1 引言

本章为解决现有融合方法未能充分利用图像深层信息的问题，延续第三章研究思路，提出了一种基于多尺度编码结构与特征增强模块的融合算法 DAE-Nest。该算法同样基于自编码器网络框架，由编码器—特征增强—融合层—解码器模块组成。首先，使用改进的编码结构提取源图像的深层多尺度特征；其次，对提取的特征进行数据增强，充分利用深层图像信息；然后，根据相应的融合策略融合增强的特征；最后，解码器接收融合特征的输入并将其重构成信息丰富的融合图像。DAE-Nest 分别在 TNO 和 RoadScene 数据集上进行了相应的融合测试实验，并与第三章 DSAFusion 以及 8 种现有的优秀融合算法进行了比较。在主观评价方面，DAE-Nest 能充分利用源图像中的深层信息，生成清晰的融合图像；在客观指标方面，DAE-Nest 在 TNO 数据集上的平均梯度（AG）和空间频率（SF）以及结构相似性（SSIM）的三个评价指标中表现最优，并在 RoadScene 数据集上的边缘信息度量（ $Q^{AB/F}$ ）评估指标中也取得了最优表现。因此，本章 DAE-Nest 在主观判断和客观评价指标上均能取得优异的融合结果，为后续的有关视觉任务奠定了良好的基础。

4.2 融合网络介绍

4.2.1 网络结构

DAE-Nest 是一个端到端的无监督融合网络，其主体结构如图 4-1 所示。其中，网络的输入是一对红外和可见光图像，分别表示为 I_{ir} 和 I_{vi} 。编码器由 5 个编码卷积块“EB”和最大池化“Pooling”交替组成。“Conv0”表示标准单位卷积。每个“EB”包含两个 3×3 卷积核的卷积层。这种纵向的编码结构能够有效地从源图像中提取更深层次的多尺度特征。

“En”表示特征增强模块，通过逐层上采样将相对深层次和相对浅层次提取的信息相结合。此外，为了防止特征冗余，“En”仅在相邻层之间执行级联操作。“En1”，“En2”，“En3”和“En4”享有同样的网络结构，但具有不同的网络参数。

“IRS”和“IRD”表示 DAE-Nest 的融合策略，与特征增强模块“En”相连。它们在不同尺度上融合增强后的红外与可见光图像特征，并将生成的融合特征传递给解码网络。

解码器部分由 6 个解码卷积块“DB”组成，采用巢形连接。“DB”和“EB”具有相同的结构，每个“DB”也包含两个卷积层。在每一行中，这些“DB”以密集连接^[38]的形式连接在一起，确保解码器能够充分利用中间层提取的信息。同时，在不同层级上，这些“DB”通过双线性插值的上采样方式从融合模块中恢复和重构多尺度特征，延续特征增强模块生成的分层信息交互特性。这些融合后的特征被收集在“DB13”中，并通过标准卷积“Conv1”输出为 I_o ，即由多尺度特征重建生成的融合图像。

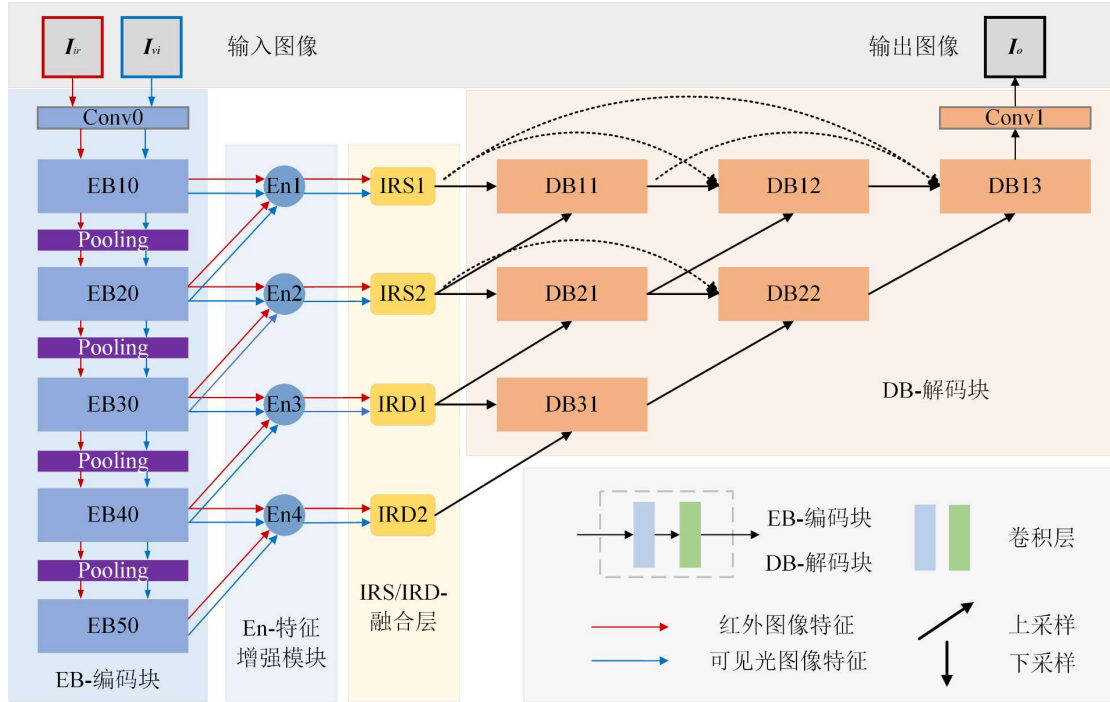


图 4-1 DAE-Nest 网络结构

具体的说，红外与可见光图像对 I_{ir} 和 I_{vi} 先被输入进了编码器中，经 5 层编码块连续下采样后产生的一系列多尺度特征 $\{f_{ir}^n\}_{n=1}^m$, $\{f_{vi}^n\}_{n=1}^m$ 。其中， n 代表编码器中特征提取层的第 n 层， m 代表总层数，此处为 5。其次，经由特征增强模块“En”对 $\{f_{ir}^n\}_{n=1}^m$, $\{f_{vi}^n\}_{n=1}^m$ 进行相应的增强后，分别在融合层进行融合得到 $\{f_F^n\}_{n=1}^5$ 。最后，经解码器作用得到融合图像。

综上，DAE-Nest 网络结构中各模块的功能为：编码器从输入的红外和可见光图像中提取多尺度特征，特征增强模块能够增强编码器提取的多尺度特征，

融合模块可以将编码器提取的来自不同模态的红外和可见光图像的多尺度特征进行融合，解码器能够将融合的特征还原重构成包含丰富信息的图像。

4.2.2 特征增强模块

DAE-Nest 的编码器从输入的源图像中提取不同尺度的信息。由于神经网络中存在深浅效应，不同尺度提取的特征会有所差异。为了更好地利用源图像的深层特征，DAE-Nest 在多尺度编码器之后连接了一个特征增强模块“En”。其具体结构如图 4-2 所示。

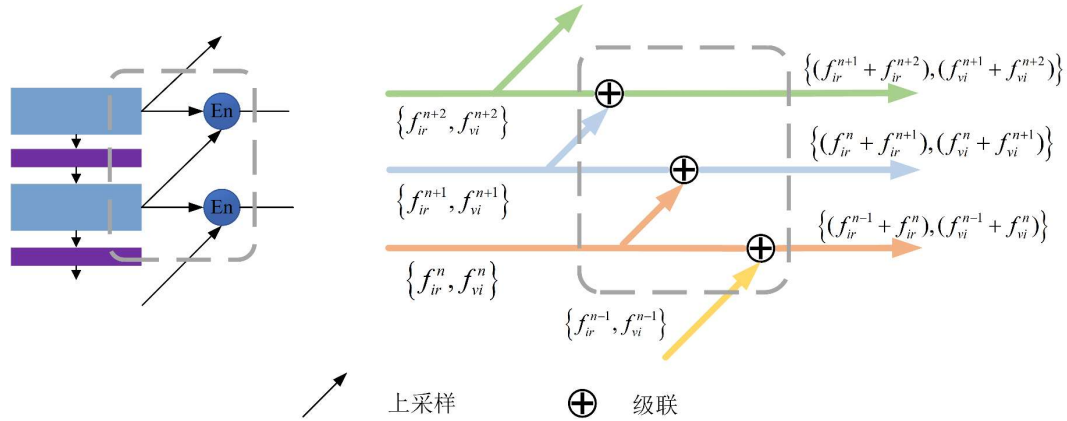


图 4-2 特征增强模块 En 网络结构

图 4-2 表示图 4-1 中“En”的内部结构，“En”先对相对深层的特征进行上采样，与相对浅层的特征进行级联，以实现每一个层级上的特征增强。需要注意的是，为了避免特征冗余，“En”仅在每相邻的两层中进行上采样与级联操作。所以从上往下是第 1、2、3、4 层级联了来自第 2、3、4、5 层的上采样信息，产生了 4 个“En”模块“En1”，“En2”，“En3”和“En4”，它们除输入输出的通道数不同，其余结构参数均相同。具体的网络参数如表 4-1 所示：

表 4-1 特征增强模块网络参数

特征增强层	卷积核	步长	输入通道数	输出通道数	激活函数
En1	1×1	1	112	64	ReLU
En2	1×1	1	160	112	ReLU
En3	1×1	1	208	160	ReLU
En4	1×1	1	256	208	ReLU

根据第 4.2.1 小节中编码器提取的多尺度特征 $\{f_{ir}^n\}$ 和 $\{f_{vi}^n\}$ ，通过“En”增强的特征可以表示为：

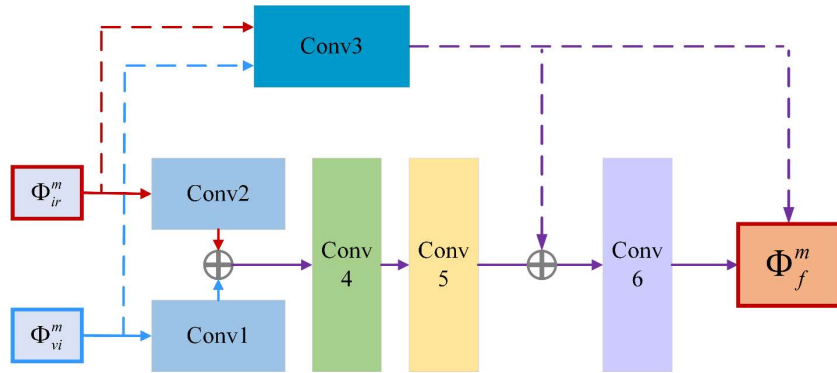
$$En\{f_{ir}^n, f_{vi}^n\} = \{(f_{ir}^n + f_{ir}^{n+1}), (f_{vi}^n + f_{vi}^{n+1})\} \quad (4-1)$$

在式（4-1）中， f_{ir}^n 和 f_{vi}^n 代表第“n”层特征提取的红外和可见光特征， f_{ir}^{n+1} 和 f_{vi}^{n+1} 代表第 $n+1$ 层提取的红外和可见光特征。变量 n 的取值范围在 1 和 4 之间。经过这样的级联操作后，由编码器提取的 5 层多尺度特征被压缩成了信息更加丰富的 4 层特征，更有利于后续的融合操作。

具体来说，DAE-Nest 的特征增强方法即利用编码器从源图像中提取的多尺度特征，将较深层尺度提取的特征，通过双线性插值的上采样方式与较浅层尺度提取的特征进行级联操作，以充分利用深层尺度所蕴含的特征信息。这样，除最底层 EB50 外，每一个与后续融合模块相连接的多尺度特征提取层均得到了不同程度的增强，并在不同的层级之间产生了信息交互。

4.2.3 融合策略

尽管现有的图像融合策略在某些应用中效果显著，但它们往往局限于简单的特征级联或仅考虑空间信息的融合方式，这些方法缺乏应用深度和端到端的处理方式，未能充分挖掘图像特征之间的互补性，限制了融合效果。DAE-Nest 考虑了多尺度特征提取存在的深浅特征蕴含信息不同的情况，并基于 Inception ResNet block^[79]的概念，设计了新的端到端的融合模块 IRS (Inception ResNet Shallow)和 IRD (Inception ResNet Deep)，具体网络结构如图 4-3 所示。



(a) IRS

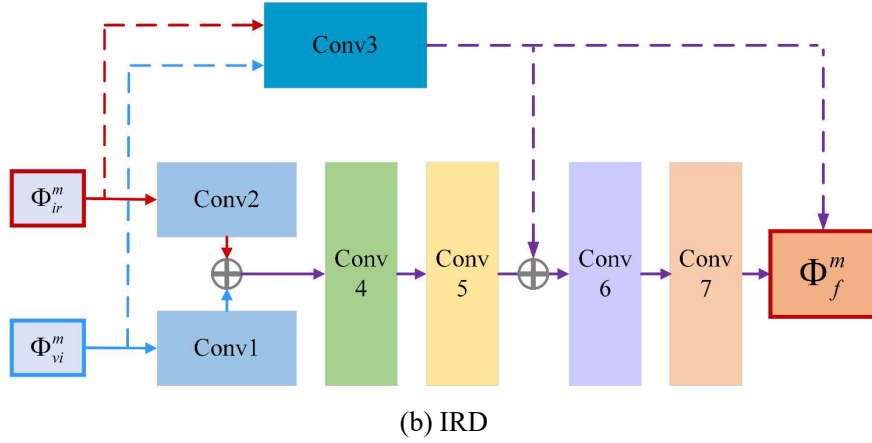


图 4-3 DAE-Nest 端到端的融合策略

图 4-3 中， Φ_{ir}^m 和 Φ_{vi}^m 分别表示红外图像和可见光图像经预训练的编码器提取出的多尺度特征， Φ_f^m 表示融合图像特征。在两种融合模块中，Conv1-Conv7 表示融合模块中的 7 个卷积层， Φ_{ir}^m 和 Φ_{vi}^m 先分别经由 3×3 的卷积 Conv1 和 Conv2，两者的输出被级联为 Conv4。同时， Φ_{ir}^m 和 Φ_{vi}^m 又被传递进了 Conv3 中，Conv3 和 Conv4 是此结构中第一个生成初始融合特征的卷积层。

根据下采样卷积的特性，浅层特征会保留更多的细节信息，而深层特征通常传递更多的语义信息。因此，IRS 的结构更简单，旨在最大化地保留来自浅层特征的纹理细节信息。IRD 在 IRS 的基础上增加了一个 3×3 的标准卷积，进一步优化了对深层特征的处理。这种改进使得 IRD 更加适合处理富含语义信息的深层特征，从而提升了特征融合的质量和效率。通过这个结构，来自不同层级的特征 Φ_{ir}^m 和 Φ_{vi}^m 可以有针对性地融合，生成不同尺度的融合特征 Φ_f^m ，这些融合特征随后被传递到解码网络，用于最终的图像重构。由于提取了多尺度深度特征和端到端的融合策略，IRS 更加专注于保留丰富的纹理细节，而 IRD 则进一步加强了对深层、含有丰富语义信息特征的处理，两者共同作用，有效地保留了红外图像中的显著目标和可见光图像中的细节纹理信息。

融合模块的具体网络参数见表 4-2。由于网络中重复利用了融合模块，存在多个输入和输出通道的组合。在此上下文中，“X”用于表示融合策略的输入和输出通道数。对应前文中提及的“IRS1”，“IRS2”，“IRD1”和“IRD2”，“X”的四个可能取值分别是 64、112、160 和 208。此外，这种设计还体现了网络对于不同输入和输出通道组合的适应性，确保了在处理多尺度特征时能够保持高效和精确，最终实现对融合图像质量的优化。

表 4-2 融合策略的网络参数

卷积层	卷积核	步长	输入通道数	输出通道数	扩展	激活函数
Conv1	3×3	1	X	X	1	ReLU
Conv2	3×3	1	X	X	1	ReLU
Conv3	3×3	1	X	X	1	ReLU
Conv4	1×1	1	X / 2	X	0	ReLU
Conv5	3×3	1	X	X	1	ReLU
Conv6	3×3	1	X / 2	X	1	ReLU
Conv7	3×3	1	X	X	1	ReLU

4.2.4 训练策略

DAE-Nest 的本质是一个端到端的融合网络，需要分步对其进行训练。同第三章 DSAFusion 一致，阶段 (I) 需要从整体网络中移除融合层，具体的网络结构如图 4-4 所示。

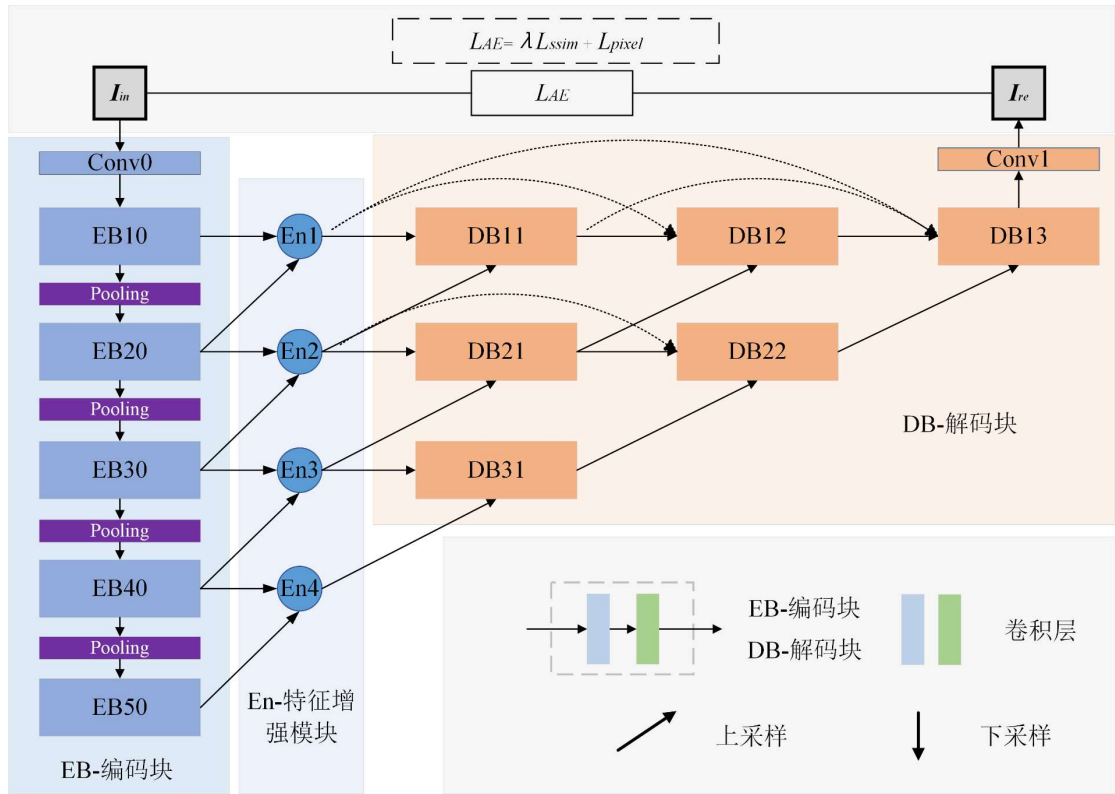


图 4-4 AE 训练过程

在图 4-4 中， I_{in} 是输入的用于训练编码器的源图像，编码器由编码块和最大池化交替构成，最大池化算子被设置成 2×2 ，它在 5 个尺度上提取深层特征。这些深层特征经特征增强模块“En”后被输入进解码网络，通过多层跳跃连接重构出的图像。同第三章训练参数一致，利用 MS COCO 数据集^[77]进行训练。

AE 的一些参数设置见表 4-3。

表 4-3 AE 训练阶段网络设置

	卷积层	卷积核	步长	输入通道数	输出通道数	激活函数
编码块	Conv0	1×1	1	1	16	ReLu
	EB10	3×3	1	16	64	ReLu
	EB20	3×3	1	64	112	ReLu
	EB30	3×3	1	112	160	ReLu
	EB40	3×3	1	160	208	ReLu
	EB50	3×3	1	208	256	ReLu
解码块	DB11	3×3	1	176	64	ReLu
	DB12	3×3	1	240	64	ReLu
	DB13	3×3	1	304	64	ReLu
	DB21	3×3	1	272	112	ReLu
	DB22	3×3	1	384	112	ReLu
	DB31	3×3	1	368	160	ReLu
	Conv1	1×1	1	64	1	ReLu

阶段 (II)训练集中的红外与可见光图像对进行融合训练，训练过程如图 4-5 所示。利用训练好的编码网络从源图像中提取多尺度深度特征，对于不同的尺度，相应的选择融合模块 IRS 或 IRD 去融合这些深度特征。然后，将融合的多尺度特征传递给对应的解码器网络。

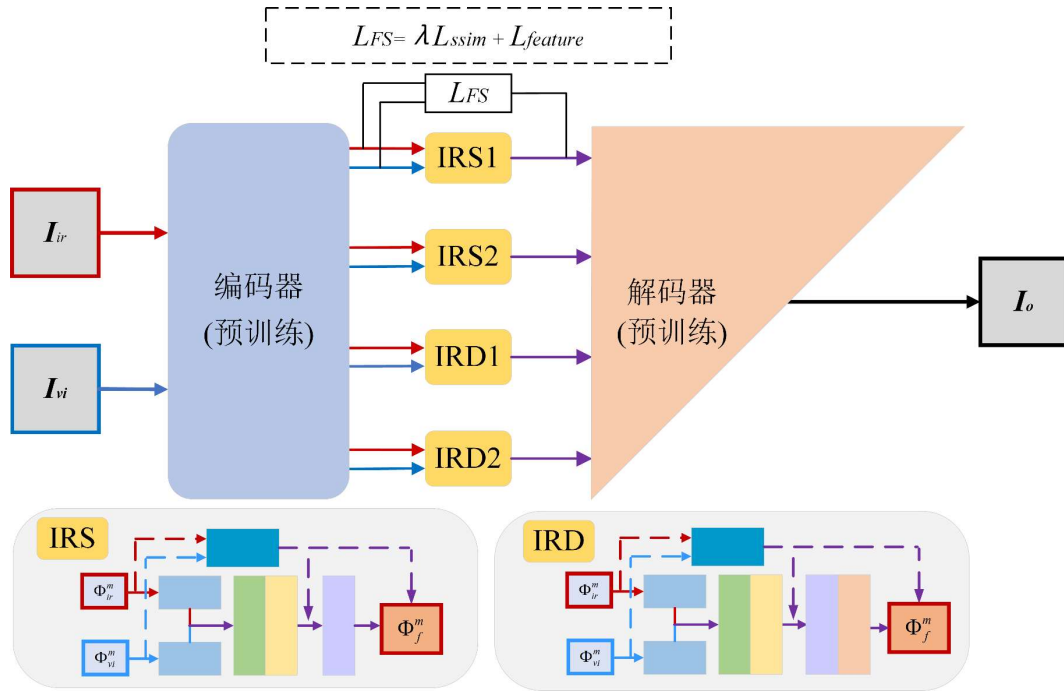


图 4-5 融合策略训练过程

阶段 (II) 融合模块的训练选择使用 KAIST 数据集^[80], KAIST 数据集包含白天和夜晚捕获的各种常规交通场景, 在该数据集中选取了 40000 对红外与可见光图像进行训练, 这些图像也被转换成灰度, 大小调整至 256×256 , 学习率 lr 为 1×10^{-4} , Batchsize 为 2, 进行 8 个 Epoch 的训练。

4.2.5 损失函数

为了获得一个能对源图像进行多尺度分解与还原重建的自编码器网络和一个能够实现端到端的融合模块的目标, DAE-Nest 使用了不同的损失函数来对两个训练过程进行约束。在阶段 (I) 训练自编码器 (AE) 时, 具体的损失函数如下:

$$L_{AE} = L_{Pixel} + \lambda L_{SSIM} \quad (4-2)$$

式 (4-2) 中, L_{AE} 、 L_{Pixel} 、 L_{SSIM} 和 λ 的定义均与 3.25 小节定义相同。 L_{Pixel} 和 L_{SSIM} 的计算公式为:

$$L_{Pixel} = ||I_{re} - I_{in}||_F^2 \quad (4-3)$$

$$L_{SSIM} = 1 - f_{SSIM}(I_{in}, I_{re}) \quad (4-4)$$

在阶段 (II) 训练融合模块时, 采用特征增强损失函数和结构相似性损失作为重构损失, 具体为:

$$L_{FS} = L_{feature} + \lambda L_{SSIM} \quad (4-5)$$

式 (4-5) 中, L_{FS} 、 $L_{feature}$ 和 L_{SSIM} 分别代表图像融合损失、特征增强损失和结构相似性损失, 为了方便计算, 这里的 λ 同式 (4-2) 取值相同, 实验设置与前文保持一致。结构相似性损失的计算方法同 (4-4), 特征增强损失的计算公式如 (4-6) 所示:

$$L_{feature} = ||\Phi_f^m - \frac{1}{2}(\Phi_{ir}^m + \Phi_{vi}^m)||_F^2 \quad (4-6)$$

综上, DAE-Nest 算法训练和测试的具体过程如图 4-6 所示, 图 4-6 中第一行代表训练阶段 (I), 通过在输入图像 I_{in} 与重建图像 I_{re} 之间建立损失 L_{AE} 来训练编码器的编码能力与解码器的还原重建能力。第二行代表训练阶段 (II) 和测试阶段, 训练阶段 (II) 利用训练阶段 (I) 训练好的自编码网络, 通过在红外图像特征 Φ_{ir}^m / 可见光图像特征 Φ_{vi}^m 与融合特征 Φ_f^m 之间建立损失 L_{FS} 来训练融合模块“IRS” / “IRD” 的特征融合能力。结束两个阶段的训练后, 即将阶段 (I) 训练好的自编码网络和阶段 (II) 的融合网络进行组合来对红外与可见光图像融合任务

进行测试。

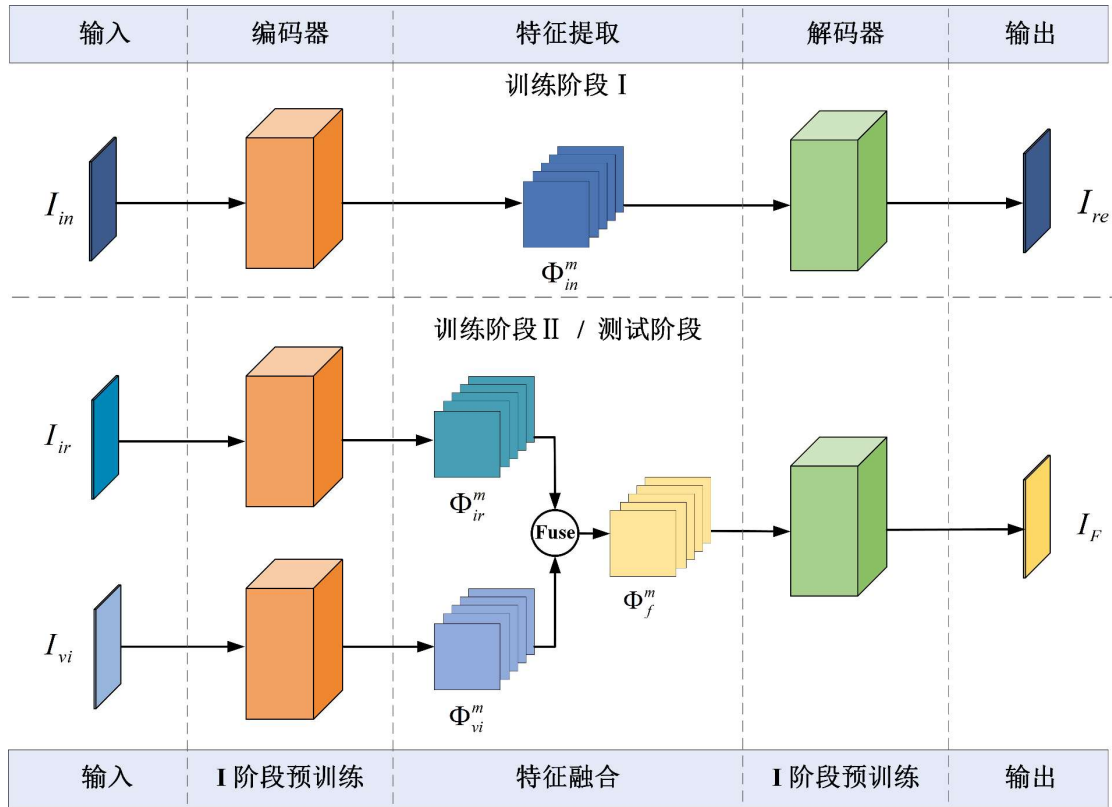


图 4-6 整体网络训练过程

4.3 融合实验及效果评估

4.3.1 实验环境及评价指标

本章 DAE-Nest 的实验环境同第三章 DSAFusion 的实验环境一致，且训练阶段与 4.2.4 节中介绍一致。测试集选用公开的红外与可见光图像数据集 TNO^[48]和 RoadScene^[50]，将数据集中的彩色图片转换成灰度图片后的样本如图 4-7 所示。

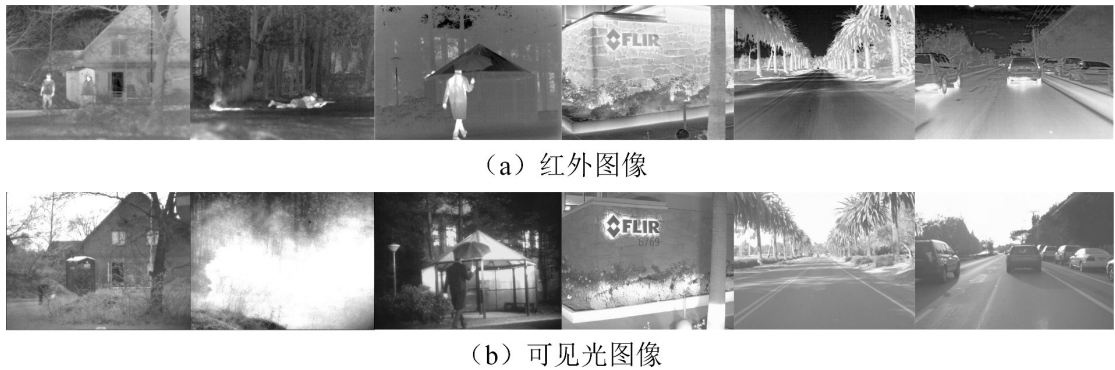


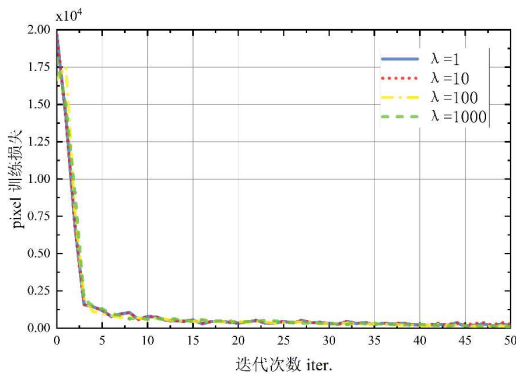
图 4-7 灰度转换后的数据集图例

对 DAE-Nest 的融合结果从主观视觉和客观指标两个方面进行对比, 选择 8 种现有的经典图像融合算法以及第三章设计的 DSAFusion 算法进行比较, 包括 RP^[81]、DCHWT^[23]、MdedFusion^[27]、DenseFuse^[38]、FusionGAN^[33]、GANMcC^[82]、NestFuse^[39]、RFN-Nest^[40]。这些融合算法的代码均可从原参考论文中获取。同时, 从前文 2.4 小节中选择了 6 个常用的评价指标来衡量融合结果, 包括 AG^[67]、MI^[71]、SF^[68]、SSIM^[70]、SD^[69]、 $Q^{AB/F}$ ^[72]。融合图像的质量均随着这 6 个评价指标数值的增加而提高。

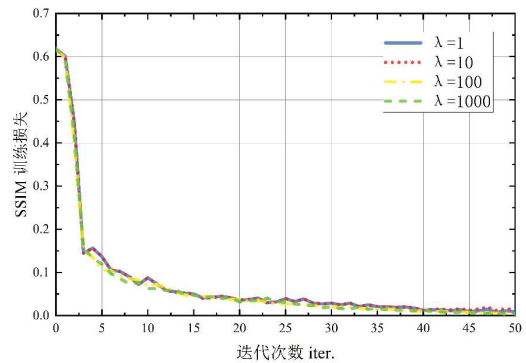
4.3.2 参数选择

4.3.2.1 平衡参数 λ 的选择

DAE-Nest 和第三章中提到的 DSAFusion 算法使用了类似的损失函数结构, 不过, DAE-Nest 采用了两阶段的训练过程, 这要求通过实验来确定损失函数中的平衡参数 λ 的最佳值。依据 4.2.5 节的讨论, 损失函数将像素损失与结构相似度损失之间的平衡参数 λ 设置为四个不同的实验值: 1、10、100, 以及 1000。在实验配置方面, 无论是编码器还是解码器的结构, 以及特征增强的策略, 都与 4.2 节中详述的设置保持一致, 保证了实验的连贯性和可比性。在融合策略上, 选择使用 IRS 来融合两层较浅尺度的特征, 采用 IRD 对两层较深尺度的特征进行融合。这种策略的选择是基于不同尺度特征对融合效果影响的理解, 旨在通过合理的层次划分来优化融合结果。以 TNO 作为测试集, 取不同的 λ 值对网络进行训练。训练损失折线图如图 4-8 所示, 其中横坐标表示迭代次数, 每个点代表 100 次迭代, 纵坐标表示训练产生的损失。



(a) 像素损失 - L_{pixel}



(b) 结构相似性损失 - L_{SSIM}

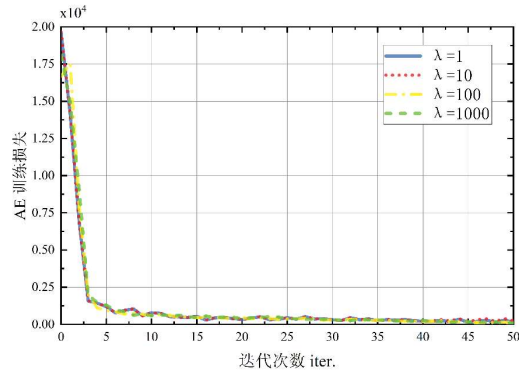
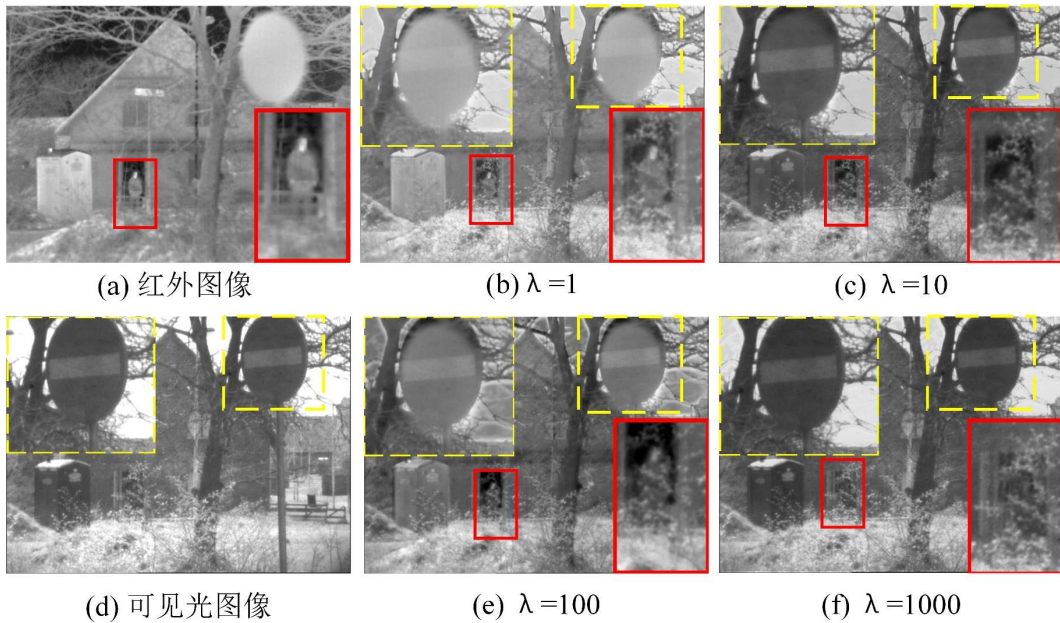
(c) 图像重建损失 - L_{AE}

图 4-8 网络训练损失

分别用“蓝”，“红”，“黄”和“绿”四种颜色代表 λ 取值为 1，10，100 和 1000 的损失情况，选择前 5000 次迭代结果绘制。由损失折线图可以看出，在前 200 次迭代中，自编码网络有着很快的收敛速度。在接近第 300 次迭代时， λ 的值取 100 或 1000 时收敛速度较快。而当迭代次数超过 400 次时，4 种取值都能得到最优的网络权重。

但总的来说，损失线图不能直观地显示不同 λ 值对网络的影响。因此，本节同时从客观评价指标方面评估了具有不同 λ 值的融合网络的性能。图 4-9 展示了不同 λ 值下的融合图像，图中红色实线框标注了红外图像中的显著目标信息，黄色虚线框标注了可见图像中的细节纹理信息。使用相应颜色的矩形框进行局部放大。

图 4-9 不同 λ 取值下的融合实验结果

从图 4-9 可以观察到, 当 λ 设为 1 时, 融合后的图像遗失了许多原始可见图像的信息, 特别是黄色虚线框内的标识显得模糊不清。相反, 将 λ 设定为 10 或 1000 导致融合图像无法有效展现显著的红外目标信息, 例如红色实线框内的人物变得难以识别。与此相对, λ 设为 100 时, 不仅显著的目标信息得到了突出显示, 而且图像的细节和纹理信息也得到了良好的保留, 实现了更好的融合效果。此外, 本小节还利用了 6 个评价指标来客观评估具有不同 λ 取值的融合网络性能。融合结果的客观指标如表 4-4 所示, 其中粗体表示最优值, 斜体表示次优值, 本章节后续所有表格均用这个标注方法表示。

表 4-4 不同平衡参数融合结果评价指标平均值

λ	AG ^[67]	MI ^[71]	SF ^[68]	SSIM ^[70]	SD ^[69]	Q ^{AB/F[72]}
1	4.305	9.509	11.37	0.471	35.41	0.379
10	5.282	11.394	14.1	0.664	40.01	<i>0.438</i>
100	<i>5.155</i>	13.784	<i>15.44</i>	0.702	63.93	0.498
1000	5.083	<i>12.399</i>	15.68	<i>0.681</i>	<i>61.14</i>	0.422

由表 4-4 可以看出, 当 λ 的值设置为 100 时, 网络获得了更优的评估指标。结合图 4-9 中 λ 在不同取值下的融合结果表现, 可以得出结论: 当 λ 设置为 100 时, 网络的整体性能最佳。因此, 在接下来的实验中, 训练阶段中损失函数的平衡参数将设置为 100。

4.3.2.2 融合策略的确定与选择

现有的一些算法已经探索了端到端的图像融合技术, 并取得了一定的成效。然而, 在这些算法的特征提取阶段, 往往忽视了图像特征在不同尺度上的多样性, 也未能针对不同尺度的特征根据其深度进行进一步的细化区分。考虑到不同深度层次的特征在融合过程中可能具有不同的重要性, 简单地使用统一的融合策略显然是不够精细的。基于此, 本研究依据前文第 4.2.3 节中介绍的融合模块结构, 对不同深度层次的特征采用不同的融合模块进行实验探索。实验使用 TNO 数据集, 并通过 6 个评价指标来考察采用不同融合模块配置的网络性能表现, 不同融合模块排列生成的融合结果如表 4-5 所示。

表 4-5 使用不同融合策略融合结果平均指标

融合策略	AG ^[67]	MI ^[71]	SF ^[68]	SSIM ^[70]	SD ^[69]	Q ^{AB/F[72]}
4IRS + 0IRD	5.967	12.221	13.24	0.639	59.36	0.411
3IRS + 1IRD	5.136	12.899	14.20	0.713	57.73	0.494
2IRS + 2IRD	5.766	13.136	14.74	0.724	60.71	0.539
1IRS + 3IRD	5.685	13.103	13.97	0.683	60.59	0.460
0IRS + 4IRD	4.076	12.981	12.44	0.646	61.14	0.508

表 4-5 中融合策略栏 $M \cdot \text{IRS} + N \cdot \text{IRD}$ 表示在特征提取过程中，从浅层到深层分别采用了“IRS”块和 N 个“IRD”块的组合，其中 M 和 N 的值都在 0 到 4 的范围内变化。根据实验结果，使用 2 个“IRS”和 2 个“IRD”块的组合在多个评价指标上表现最好，这一发现也从侧面证实了所提出的融合方法的有效性。因此，在 DAE-Nest 模型中，选用两个“IRS”作为浅层的融合策略，两个“IRD”作为深层的融合策略。

此外，为进一步探究“IRD”相对于“IRS”的改进效果，并研究不同卷积核尺寸的“Conv7”在深层特征融合中的适用性，本小节设计了一个实验比较了无“Conv7”卷积层的“IRD”（即相当于“IRS”）与包含不同卷积核尺寸的“Conv7”（分别为“ 1×1 ”、“ 3×3 ”和“ 5×5 ”）的效果。通过采用一系列客观评价指标来评估各配置下的融合效果，以此验证不同卷积核尺寸的“Conv7”在深层融合中的有效性。在融合网络结构设计中，前两层较浅的融合策略仍采用“IRS”结构，而在后两层较深的部分，则选用了配备有不同“Conv7”卷积核的“IRD”结构作为融合策略，具体的实验结果如表 4-6 所示。

表 4-6 “IRD”中不同卷积核大小的“Conv7”的实验结果

“Conv7”的 卷积核尺寸	AG ^[67]	MI ^[71]	SF ^[68]	SSIM ^[70]	SD ^[69]	Q ^{AB/F[72]}
无	5.967	12.221	13.24	0.639	59.36	0.411
1x1	5.422	12.452	13.18	0.654	60.20	0.487
3x3	5.766	13.136	14.74	0.724	60.71	0.539
5x5	4.489	12.527	11.74	0.514	50.69	0.429

根据表 4-6 中的数据，显而易见地，当融合模块“IRD”中的卷积核“Conv7”设为“ 3×3 ”时，整体融合性能得到了增强。该设置优于其它方法，为采用不同融合模块处理不同尺度提供了间接证据。

4.3.3 实验结果分析

将本章 DAE-Nest 算法与第 4.3.1 小节提到的现有融合方法进行比较。不同方法的融合结果比较如图 4-10 和 4-11 以及图 4-12 和 4-13 所示。图 4-10 和 4-11 中的实验结果取自 TNO^[48]数据集，图 4-12 和图 4-13 中的实验结果取自 RoadScene^[50]数据集。图中的红色实线框表示红外图像中的显著目标信息，黄色虚线框表示可见图像中的详细纹理信息。同时使用相应颜色的矩形框进行局部放大。

从图 4-10 和 4-11 中可以看出，RP^[81]算法采用了像素级别的图像融合技术，尽管这种方法在理论上能够实现细节信息的高度保留，但实际融合结果受到了噪声的明显影响。这种影响导致在两组对照图像中出现了大量细节信息的丢失，反映出该算法在噪声抑制和细节信息保留方面的不足。与 RP^[81]算法相比，DCHWT^[23]和 MdFusion^[27]算法显示出优异的性能，特别是在抑制噪声和保留图像纹理细节方面。这两种算法通过更加精细的处理方式，有效地避免了噪声的引入，从而在纹理信息的保留上取得了显著成效。然而，尽管在这些方面表现突出，它们仍旧未能在融合图像中充分突出显著的目标信息。进一步地，FusionGAN^[33]、NestFuse^[39]和 RFN-Nest^[40]算法整体上实现了较为理想的融合效果。这些算法不仅有效地维持了图像的总体质量，还在一定程度上解决了细节信息保留的问题。尽管如此，FusionGAN^[33]和 NestFuse^[39]在处理可见光图像的细节纹理信息方面仍显不足，如图 4-10 所示，黄色虚线框中的门牌信息在融合后变得模糊，说明这些算法在细节纹理的捕捉和保留方面还有提升空间。相比之下，RFN-Nest^[40]虽然在维持详细信息方面做得较好，但在强调显著目标信息上的能力仍有限。在 GANMcC^[82]和 DenseFuse^[38]算法的对比中，两者均展现了在保留纹理细节方面的强大能力。不过，GANMcC^[82]算法在某些情况下会生成较暗的图像，例如在图 4-11 中红色框下方的场景信息因过度暗淡而丢失，这可能会影响到融合图像的整体观感和信息的完整性。DenseFuse^[38]算法虽然避免了图像过暗的问题，但其融合结果却偏向于过亮，同时红外显著目标的轮廓也出现了模糊，这减少了图像的清晰度和可辨识度。DSAFusion 算法在生成的图像清晰度方面表现较好，但在详细信息和纹理保留方面，仍有进一步优化的空间。相较而言，本章提出的 DAE-Nest 算法在多个方面都显示出了卓越的性能。它不

仅保留了可见光图像中的细节纹理信息，如图 4-10 和 4-11 中黄色虚线框内店铺信息和门框的清晰展现所证明，而且在显著目标的突出上也做得非常成功，生成的融合图像具有更高的对比度和清晰度，更好地适应了人类视觉感知系统。

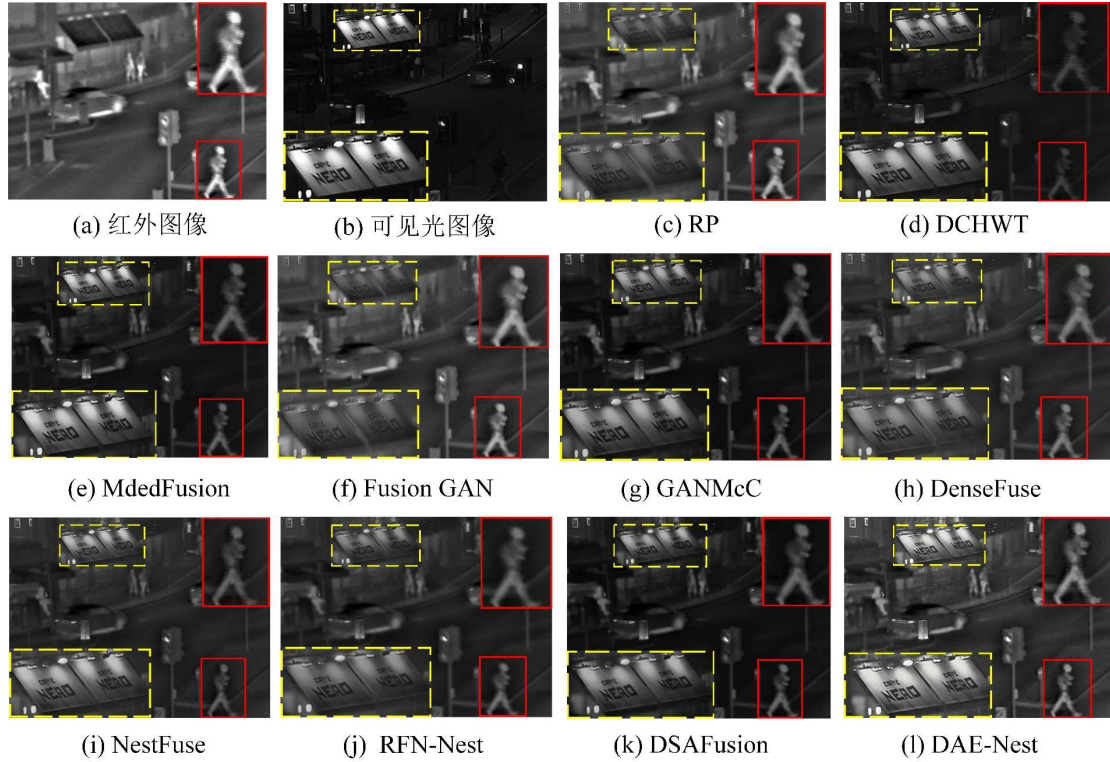


图 4-10 不同算法在 TNO 测试集上的融合结果一

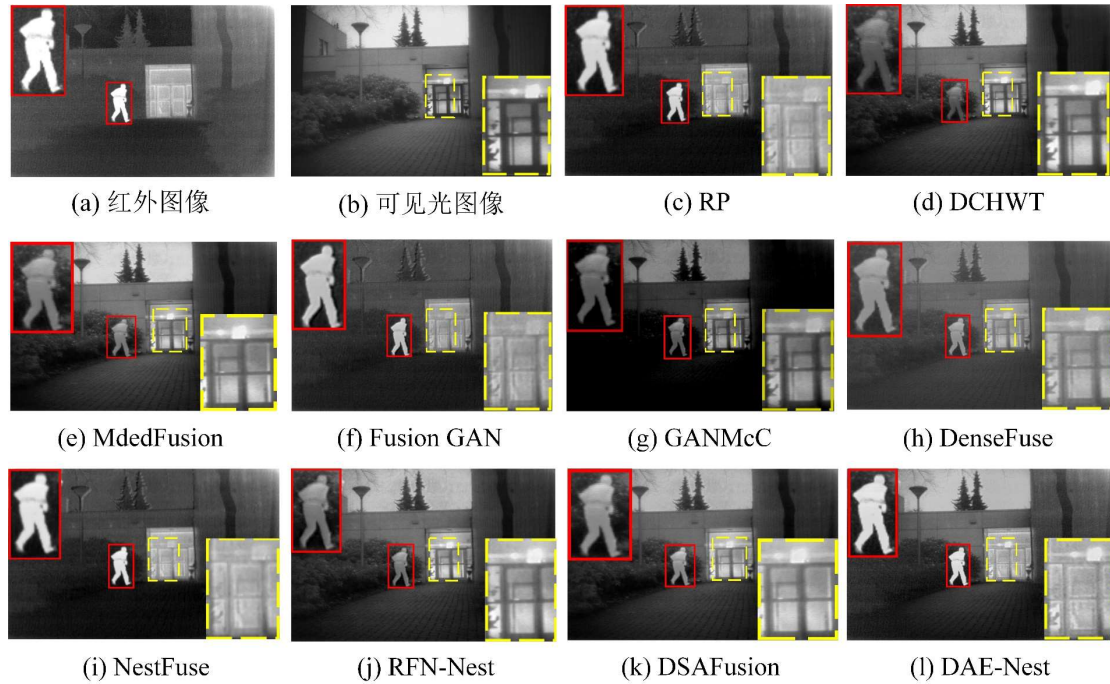


图 4-11 不同算法在 TNO 测试集上的融合结果二

与 TNO 测试集的专门场景不同, RoadScene 测试集提供了一个更加广泛和多样化的户外道路驾驶环境, 涵盖了从城市街道到郊外道路的各种场景。这些场景中的图像往往具有较低的清晰度和较为复杂的背景, 包括但不限于各种交通工具、路标、行人以及自然和人造结构, 有利于测试和验证图像融合算法在实际应用中的效果, 尤其是在处理模糊图像和背景杂乱的条件下的性能。如图 4-12 和图 4-13 所示, 不同的图像融合算法在这一测试集上的表现存在差异。RP^[81]算法的融合结果表明, 在处理模糊和复杂场景时, 其融合图像的清晰度明显下降, 结果严重模糊, 几乎无法识别关键的场景元素, 如车辆和路标等。DCHWT^[23]和 FusionGAN^[33]算法虽然在一定程度上保留了背景的纹理信息, 使得融合图像在视觉上更为接近真实场景, 但这两种算法未能有效捕捉和强调红外图像中的显著目标信息。这意味着在关键信息的捕捉上, 这些算法未能充分利用红外图像的优势, 从而减少了融合图像在特定应用场景下的价值。另一方面, MdedFusion^[27]、GANMcC^[82]、DenseFuse^[38]和 NestFuse^[39]算法虽然没有产生过多噪声, 显示出一定程度的稳定性, 但详细信息的严重丢失问题依然存在。在图 4-12 和图 4-13 中的黄色虚线框示例中, 车辆和路标这些关键元素难以被清晰辨识。在此背景下, RFN-Nest^[40]、DSAFusion 和 DAE-Nest 算法通常能够提供更优的融合效果, 显示出在维持图像清晰度和强调关键细节方面的优势。特别是, DAE-Nest 算法在保持整体图像质量的同时, 能够有效突出显著的目标信息, 例如在图 4-13 中红色实线框内行人信息的显著性明显优于其它算法, 展现出了对道路细节的优良强调能力。相比之下, DSAFusion 算法虽然在综合融合效果上表现不错, 但其生成的图像在对比度方面稍显不足, 可能影响到图像的视觉效果。而 RFN-Nest^[40]算法生成的融合图像虽整体较暗, 但在道路细节的强调方面略显不足, 尤其是在突出行人等关键信息的表现上不如 DAE-Nest 算法。

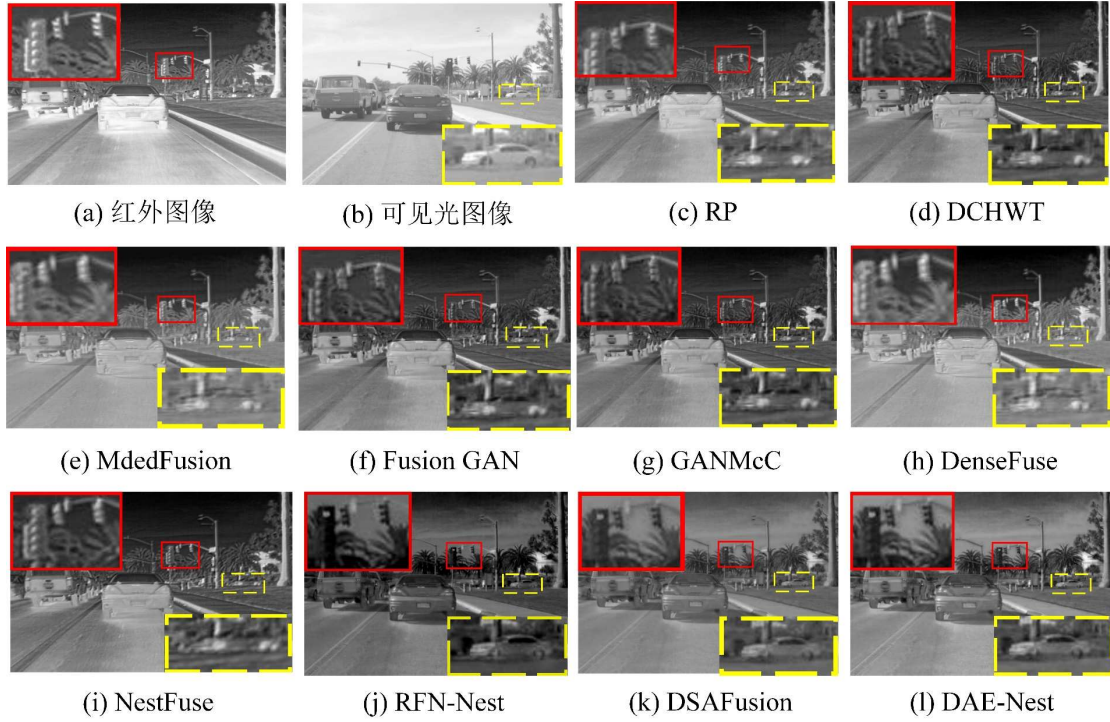


图 4-12 不同算法在 RoadScene 测试集上的融合结果一

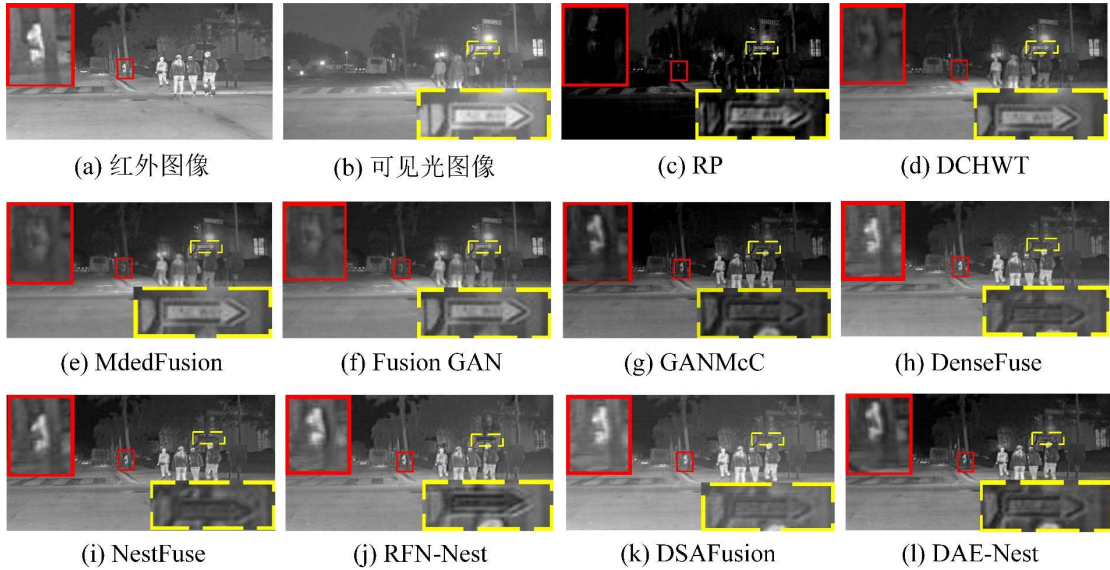


图 4-13 不同算法在 RoadScene 测试集上的融合结果二

总的来说，其它融合方法存在纹理细节被噪声破坏或者显著目标被削弱的问题。相比之下，DAE-Nest 能够有效地对红外图像中的显著目标信息和可见光图像中的纹理细节信息进行融合。

除了主观视觉判断，本章同第三章一致，采用了 6 个客观评价指标来比较这些融合方法。仍然使用 TNO 和 RoadScene 数据集进行实验。TNO 数据集上的实验结果如表 4-7 所示，RoadScene 数据集上的结果如表 4-8 所示。

表 4-7 不同算法在 TNO 测试集上的评价指标平均值

融合方法	AG ^[67]	MI ^[71]	SF ^[68]	SSIM ^[70]	SD ^[69]	Q ^{AB/F[72]}
RP ^[81]	4.879	12.835	6.88	0.625	39.56	0.322
DCHWT ^[23]	4.993	12.988	7.02	0.653	41.37	0.379
MdedFusion ^[27]	5.356	13.298	9.49	0.681	52.56	0.412
FusionGAN ^[33]	5.166	13.472	6.91	0.676	50.16	0.369
GANMcC ^[82]	5.783	13.087	8.39	0.683	55.23	0.439
DenseFuse ^[38]	5.476	12.890	7.44	0.704	46.14	0.383
NestFuse ^[39]	5.664	13.156	8.98	0.727	49.36	0.412
RFN-Nest ^[40]	5.758	13.388	9.33	0.735	52.10	0.488
DSAFusion	5.569	13.633	9.27	0.716	60.83	0.426
DAE-Nest	5.983	13.346	9.75	0.748	56.34	0.467

表 4-8 不同算法在 RoadScene 测试集上的评价指标平均值

融合方法	AG ^[67]	MI ^[71]	SF ^[68]	SSIM ^[70]	SD ^[69]	Q ^{AB/F[72]}
RP ^[81]	3.677	9.221	8.24	0.675	40.37	0.318
DCHWT ^[23]	4.136	10.899	8.20	0.713	45.73	0.392
MdedFusion ^[27]	6.033	11.596	13.94	0.759	54.33	0.462
FusionGAN ^[33]	4.766	10.136	12.74	0.742	57.69	0.456
GANMcC ^[82]	4.685	11.103	12.97	0.827	55.48	0.423
DenseFuse ^[38]	4.076	11.981	10.44	0.746	57.44	0.396
NestFuse ^[39]	5.488	12.498	12.87	0.785	56.98	0.462
RFN-Nest ^[40]	6.286	13.673	13.56	0.833	57.05	0.478
DSAFusion	6.196	13.569	13.59	0.826	61.64	0.433
DAE-Nest	6.739	12.780	14.43	0.861	58.84	0.513

从表 4-7 可以看出，与 TNO 数据集融合结果的其它融合算法相比，本章 DAE-Nest 算法在平均梯度（AG^[67]）和空间频率（SF^[68]）的最优值表明所提出的融合方法可以从源图像中提取更多信息。结构相似性（SSIM^[70]）的最优值也显示出所提融合方法可以保留更多特征信息。表 4-8 中的实验结果表明，本章 DAE-Nest 在不同数据集上的实验结果基本相同，在 RoadScene 数据集上，DAE-Nest 不仅保留了最佳的平均梯度（AG^[67]）、空间频率（SF^[68]）和结构相似性（SSIM^[70]），还在边缘信息度量（Q^{AB/F[72]}）评估指标上取得了最佳值。这也表明该算法具有良好的泛化能力。

4.3.4 融合结果检测实验

此外，为了验证 DAE-Nest 是否可以提高后续视觉任务的准确性，将来自 TNO^[48]数据集的红外图像、可见光图像以及 MdedFusion^[27]、GANMcC^[82]、DSAFusion 和 DAE-Nest 等算法生成的融合图像输入到 YOLOv5^[83]检测网络中。单目标场景的检测结果如图 4-14 所示：

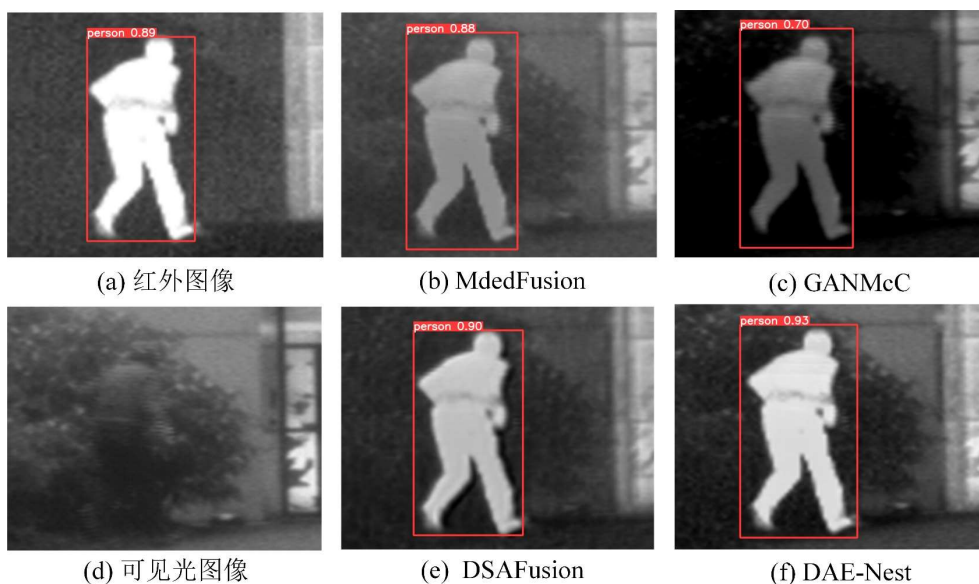


图 4-14 单目标场景下 YOLOv5 检测结果

图 4-14 显示了单目标场景下原始图像和不同融合方法在 YOLOv5 网络中的检测结果。其中，红外图像的检测结果为 0.89，对于 MdedFusion^[27]、GANMcC^[82]、DSAFusion 和 DAE-Nest 等融合算法的检测结果分别为 0.88、0.70、0.90 和 0.93。由于可见光图像提供的显著性信息不足，YOLOv5 网络未能实现有效检测。图 4-15 展示了多目标场景下的检测结果，但由于多目标场景下检测数据众多，识别框产生了重叠，为了更直观的展示检测结果，将图 4-15 的检测数据誊写在表 4-9 中展示：

表 4-9 多目标场景下的检测结果

目标	红外图像	MdedFusion ^[27]	GANMcC ^[82]	可见光图像	DSAFusion	DAE-Nest
1	0.58	0.68	0.76	0.69	0.82	0.87
2	0.64	0.70	0.72	0.68	0.86	0.91
3	0.42	0.54	0.63	0.53	0.79	0.83

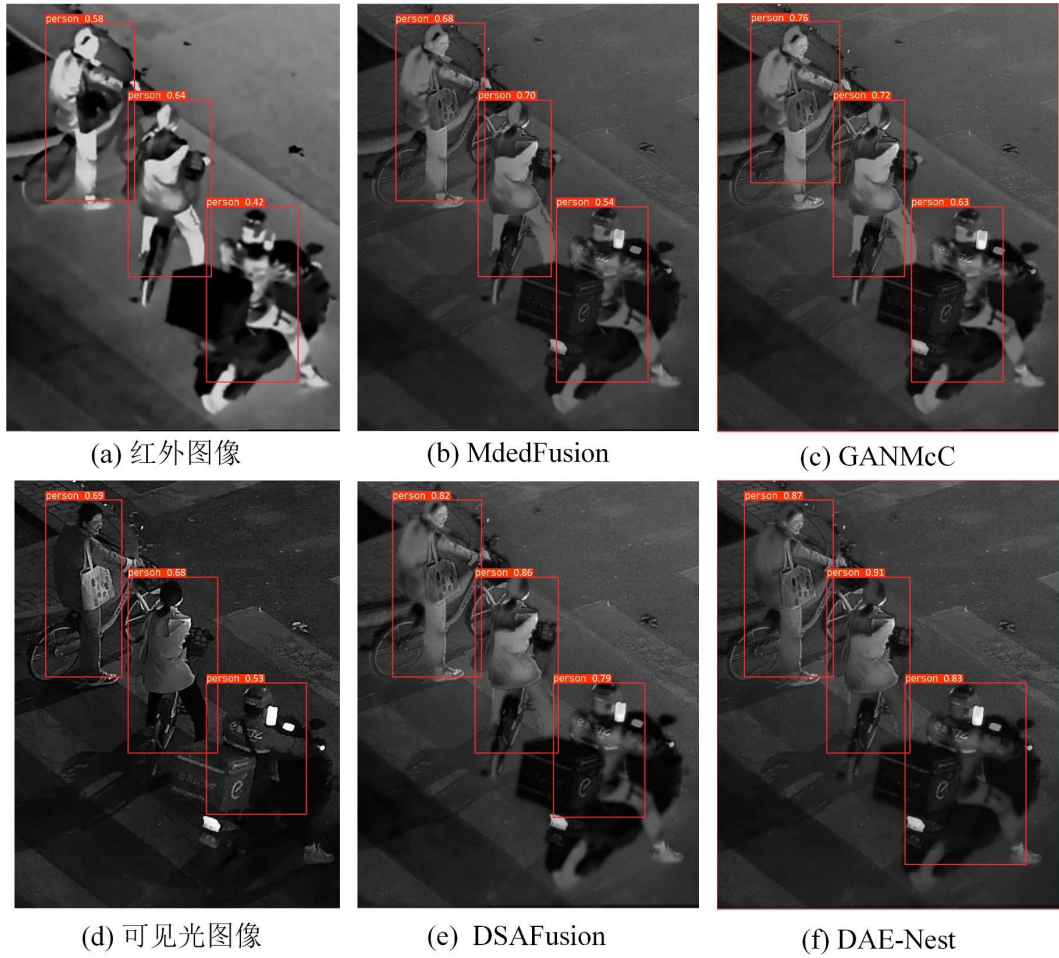


图 4-15 多目标场景下 YOLOv5 检测结果

由图 4-14、4-15 和表 4-9 可以看出，本章 DAE-Nest 与第三章 DSAFusion 均能成功融合不同模态的图像以增强图像质量，有利于后续的有关视觉任务。此外，在不同融合方法的检测结果中，DAE-Nest 也取得了最高的准确率，进一步证明了本章所提方法的出色融合性能。

总的来说，DAE-Nest 有着合理的损失函数和特征增强模块，并充分利用了编码器提取的深层多尺度特征，在保留源图像的细节纹理信息和增强显著目标信息方面具有优势。关于 DAE-Nest 生成的融合结果，无论是从主观判断还是客观数据方面，该算法生成的图像都具有更好的清晰度和更强的对比度，符合人的视觉感知系统，并且在红外和可见图像信息的平衡方面取得了良好的效果。

4.3.5 消融实验

为了验证 DAE-Nest 中深层多尺度特征提取和特征增强模块的作用，本小节采用 21 对来自 TNO 数据集中的红外和可见光图像进行消融实验。并对各种实

验生成的融合图像进行了客观评估。具体而言，对于编码器中的深层多尺度特征提取，通过改变多尺度特征提取层中的层数为 4、5 和 6 来进行消融实验；而对于特征增强模块“En”，则直接观察它的加入对融合结果产生的影响。在此实验中，除了编码器中多尺度特征提取层的层数和是否包含特征增强模块“En”之外，其它配置如学习率、训练轮数、损失函数平衡参数等都保持固定并与前文一致。实验结果如表 4-10 所示，“Exp”表示实验编号，“Scale layers”表示多尺度特征提取中的层数，“En”表示是否使用第 3.2 节中提出的特征增强模块。

表 4-10 特征提取尺度和特征增强模块消融实验

Exp	Scale layers	En	AG ^[67]	MI ^[71]	SF ^[68]	SSIM ^[70]	SD ^[69]	Q ^{AB/F} ^[72]
1	4	--	4.833	11.324	8.23	0.635	51.09	0.434
2	4	√	5.132	11.965	9.27	0.715	53.22	0.446
3	5	--	5.352	11.689	9.93	0.698	53.76	0.419
4	5	√	5.983	13.346	9.75	0.748	56.34	0.467
5	6	--	4.566	10.980	7.37	0.576	48.67	0.385
6	6	√	4.713	11.673	7.89	0.537	49.03	0.428

表 4-10 呈现了使用不同模块组合的实验评估指标，从中可以看出，Exp.1 和 Exp.2 的特征提取层层数设置为 4，Exp.3 和 Exp.4 的特征提取层层数设置为 5，而 Exp.5 和 Exp.6 的特征提取层层数设置为 6。当多尺度特征提取层的数量设置为 5 时，实验数据优于 4 和 6。这可能因为当特征提取层的层数为 4 时，对图像的多尺度特征提取不完整，而当把层数设置为 6 时，提取的特征过多导致融合图像出现了特征冗余。

Exp.2、Exp.4 和 Exp.6 包含了本章 4.2.2 节中提出的特征增强模块“En”，而 Exp.1、Exp.3 和 Exp.5 则没有。相应的实验数据表明，特征增强模块“En”的加入对融合结果有积极的增强作用。它有效提高了图像融合的质量，减少了图像的噪声或失真，赋予了融合图像更多的信息、可读性和适应性。充分证明了 DAE-Nest 中特征增强模块“En”的有效性。

4.3.6 网络融合效率分析

除此之外，为了验证 DAE-Nest 的融合效率，本小节进行了一项测试实验，使用来自 TNO^[48]和 RoadSence^[50]数据集的 40 对随机选择的红外和可见光图像，融合算法的效率是基于对这 40 对图像融合所需的平均时间衡量。本小节将

DAE-Nest 的平均运行时间与其它网络进行了比较, 这些网络包括 DCHWT^[23]、GANMcC^[82]、RFN-Nest^[40]以及第三章设计的 DSAFusion, 它们在融合结果上都具有较优的表现。值得注意的是该实验与 3.3.4 小节中测量 DSAFusion 算法运行时间的实验配置、步骤相同, 但图像对是随机选取的, 两次测量结果会有细微差异, 具体实验结果如表 4-11 所示。

表 4-11 不同算法对 40 个图像对的平均处理时间

融合算法	DCHWT ^[23]	GANMcC ^[82]	RFN-Nest ^[40]	DSAFusion	DAE-Nest
平均运行时间 (s)	1.228	0.917	0.359	0.184	0.312

由表 4-11 的数据可以明显看出, 得益于简单的网络结构以及八度卷积处理能力, DSAFusion 有着最快的网络运行速度, 仅为 0.184 秒。而 DAE-Nest 相较于其它融合算法也表现出了较快的融合速度, 为 0.312 秒。结合表 4-7 和表 4-8 中的评价指标与表 4-10 中的运行时间信息, 可以得出结论: DAE-Nest 算法不仅有效提升了红外和可见光图像融合的质量, 而且还适当加快了网络的运行速度, 进一步提高了图像融合的效率。

4.4 DSAFusion 和 DAE-Nest 算法对比

本研究中第 3 章与第 4 章中所设计的 DSAFusion 和 DAE-Nest 都是针对红外与可见光图像的不同融合算法, 两者具有很多的共性与特性。因此, 结合前文内容, 本小节将从网络结构设计 with 实验数据比对等多个方面, 深入讨论两种方法的联系和差异性。

(1) 特征提取结构

DSAFusion 针对现有融合网络常从单一尺度提取特征的局限, 结合密集链接模块、语义信息模块设计了基于双尺度的编码结构, 以更全面的提取图像特征。而 DAE-Nest 延续 DSAFusion 的设计思路, 在双尺度特征提取的基础上, 利用快速下采样将输入的源图像分成五个层级, 再利用特征增强模块对这五个层级进行强化, 从而充分利用深层信息, 获得源图像更全面的特征表示。

(2) 网络训练方式

DSAFusion 与 DAE-Nest 算法同属于自编码器网络框架, 但不同的是 DSAFusion 利用通道注意力机制和空间注意力机制作为融合策略, 仅需对编码、解码过程进行训练。而 DAE-Nest 使用了新设计的端到端的融合模块“IRS”和

“IRD”，这两种融合模块根据特征的尺度进行融合，极大程度的保留了来自图像的深层特征信息。因此，DAE-Nest 需要分步训练，除了对编码、解码环节进行训练外，还需要额外训练端到端的融合网络。

（3）融合效果分析

从 4.3.3 小节中图 4-10、图 4-11、图 4-12 以及图 4-13 中可以看出，两种方法生成的融合图像风格相似，但 DAE-Nest 在凸显红外显著目标方面比 DSAFusion 略好。结合表 4-7、表 4-8 数据可以看出，DAE-Nest 在 TNO、RoadScene 数据集上的测试结果分别取得了 3 项、4 项最优，而 DSAFusion 则分别取得了 2 项、1 项最优，DAE-Nest 获得了更多评价指标的最优值，这也意味着其相较于 DSAFusion 有着更好的融合效果。

（4）检测结果分析

从 4.3.4 小节中图 4-14 以及图 4-15 可以看出，DSAFusion 与 DAE-Nest 均可有效的完成图像融合，以促进后续的视觉检测任务。其中，DSAFusion 在单目标场景下对人物的检测结果为 0.90，多目标场景下对多个目标的检测结果分别是 0.82、0.86 和 0.79，而在同一实验中，DAE-Nest 对单目标的检测结果为 0.93，多目标的检测结果分别为 0.87、0.9 和 0.83。由此可见，得益于更细致的特征处理，DAE-Nest 对融合后续任务的处理能力要优于 DSAFusion。

（5）融合效率分析

从 4.3.6 小节中表 4-11 对多种融合算法的平均运行时间对比中可以看出，DSAFusion 与 DAE-Nest 处理 40 个图像对的平均运行时间分别为 0.162 和 0.312 秒，证明了 DSAFusion 的融合效率要高于 DAE-Nest。

综上所述，DSAFusion 和 DAE-Nest 算法均能有效完成红外与可见光图像融合任务以及利于后续的视觉任务。不同的是，DSAFusion 着重于网络运行效率，整体网络结构较为简单，而 DAE-Nest 着重于更好的图像融合效果，在训练过程中会消耗更多的算力。

4.5 本章小结

本章延续第 3 章的设计思路，针对源图像中的深层特征进行多尺度提取，提出了一种基于深度特征提取与增强的融合算法 DAE-Nest。首先，利用编码网络对源图像进行深层的特征提取，从而获得了来自源图像的不同尺度的特征；

接着将相应深层的特征逐层加权到相应浅层的特征上，以实现对特征的增强；然后把增强后的特征根据包含信息的侧重选择对应的融合策略进行融合；最后由解码器将这些特征重构成一幅融合图像。DAE-Nest 同样在 GPU 上基于 PyTorch 框架实现，并在不同的数据集上与不同现有融合方法进行了比较，均取得了较为优秀的融合结果，说明 DAE-Nest 可以充分提取源图像中的特征信息，并且保证了融合图像有着较高的清晰度，更加符合人眼的视觉感知。

第5章 总结与展望

5.1 总结

红外与可见光图像融合技术旨在整合红外图像与可见光图像中的补充信息，创建信息丰富的融合图像，以实现对观测场景更全面、精准的描述，有利于后续视觉任务。随着该领域概念的不断发展，众多融合技术相继被开发，取得了显著的成果。尽管如此，现有方法仍有改进空间。本文对现有的图像融合技术展开了广泛的分析，针对其中的缺陷与不足，提出了两种基于改进自编码器的红外与可见光图像融合算法。并通过系列实验，验证了所提算法的有效性。主要研究内容包括：

(1) 为应对红外特征与可见光特征在融合图像中的表征失衡问题，本文提出了一种基于双尺度编码结构与注意力机制的红外与可见光图像融合算法 **DSAFusion**。首先，通过引入密集链接模块和语义信息模块构建的双尺度特征提取网络，有效提取源图像的细节和语义信息，为图像融合提供丰富特征；其次，引入 **CoTNet** 注意力机制强化信息表达，优化融合图像质量；接着，在融合策略中加入空间和通道注意力机制，最大化保留重要特征；最后，通过解码器还原重建，生成高质量的融合图像。除此之外，**DSAFusion** 通过利用八度卷积替换标准卷积，在保证融合质量的同时有效提升了网络运行效率。

(2) 针对源图像中深层信息无法得到充分利用的局限，本文提出了一种基于多尺度编码结构与特征增强模块的红外与可见光图像融合算法 **DAE-Nest**。具体的，首先通过连续下采样捕获源图像不同尺度的特征，充分理解图像的全局与局部信息；随后，利用特征增强模块“**En**”对多尺度特征进行逐级增强，提升特征表达能力，确保充分利用图像信息；接着，增强后的特征将被输送进融合模块，该模块根据特征的尺度深浅情况，通过“**IRS**”和“**IRD**”组合进行融合，确保了不同尺度特征的有效结合；最后，采用巢型连接的解码器逐层还原重建融合特征成最终图像，保留融合细节，确保最终生成的图像质量。

(3) 为确保 **DSAFusion** 与 **DAE-Nest** 算法的融合效果，本文针对不同算法的融合过程定制了不同的训练计划，并用相应实验进行了融合质量评估。对于 **DSAFusion** 算法，直接去除融合网络部分，单独对编码器、解码器进行训练。

而对于 DAE-Nest 算法,除了训练编码、解码环节外,还需要额外训练端到端的融合网络。利用合理设计的损失函数和训练策略,两种算法均能产生更丰富信息和更高质量的融合图像。通过在不同数据集上分别进行测试,试验结果表明,DSAFusion 和 DAE-Nest 融合算法均能有效的完成红外与可见光图像的融合,这两种融合技术不仅提升了图像的视觉效果,而且为后续的视觉任务如目标检测、场景理解等,提供了有利的支持。

5.2 展望

本文通过对现有的红外与可见光图像融合算法展开广泛的分析后,提出了上述两种基于改进自编码器的算法。实验结果表明,这两种算法具有较好的融合效果,但仍存在许多不足之处,在未来工作中还可进一步改进和完善,具体包括以下几个方面:

(1) 为了保证融合图像的效率,DSAFusion 融合算法尚未探讨更多尺度模块以及其它模块的使用是否能带来更好的融合效果。后续应侧重深入探索多尺度模块的应用潜力、优化网络结构以提升融合效果与效率、改善训练阶段的损失函数以减少参数计算量等方面,并应在更广泛的数据集上进行测试以增强融合模型的泛化能力。这些方面的深入研究旨在进一步提高融合图像的质量,降低计算成本,并确保算法能够在不同的实际应用场景发挥较好的性能。

(2) 为了保证融合图像的质量,DAE-Nest 融合算法使用了两阶段的训练策略,有效的增加了资源的利用率,但同时其复杂的网络结构也使得对融合质量的优化带来了不确定性。针对这些挑战,未来的研究方向将包括简化网络结构以减轻计算负担和资源消耗,并探索多任务学习的可能性,提升模型的泛化能力。这些深入探索的目的是开发出既高效又灵活的图像融合算法,能够在保持或提升融合图像质量的同时,降低对计算资源的需求,并适应多种图像处理任务,从而在实际应用中具有更广泛的适用性和更好的性能表现。

参考文献

- [1] Wang Z, Ziou D, Armenakis C, et al. A comparative analysis of image fusion methods[J]. IEEE transactions on geoscience and remote sensing, 2005, 43(6): 1391-1402.
- [2] White F E. Data fusion lexicon[J]. Joint Directors of Laboratories, Technical Panel for C, 1991, 3: 19.
- [3] 王丽芳, 史超宇, 蔺素珍, 等. 基于联合图像块聚类自适应字典学习的多模态医学图像融合[J]. 计算机科学, 2019, 046(007) : 238-245.
- [4] 徐东东.基于无监督深度学习的红外与可见光图像融合方法研究[D].中国科学院大学(中国科学院长春光学精密机械与物理研究所),2020.
- [5] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.
- [6] Haghighat M, Razian M A. Fast-FMI: Non-reference image fusion metric[C]//2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT). IEEE, 2014: 1-3.
- [7] Tang H, Xiao B, Li W, et al. Pixel convolutional neural network for multi-focus image fusion[J]. Information Sciences, 2018, 433: 125-141.
- [8] Su W, Huang Y, Li Q, et al. GeFuNet: A knowledge-guided deep network for the infrared and visible image fusion[J]. Infrared Physics & Technology, 2022, 127: 104417.
- [9] Chen S, Zhang R, Su H, et al. SAR and multispectral image fusion using generalized IHS transform based on à trous wavelet and EMD decompositions[J]. IEEE Sensors Journal, 2010, 10(3): 737-745.
- [10] Shreyamsha Kumar B K. Image fusion based on pixel significance using cross bilateral filter[J]. Signal, image and video processing, 2015, 9: 1193-1204.
- [11] 柯聪.基于改进生成对抗网络的红外与可见光图像融合算法研究[D].湖北工业大学,2021.
- [12] Choi Y K, Pak J H, Seo K, et al. Realization of infrared display images using infrared laser projection method[J]. Optics and Lasers in Engineering, 2021, 145: 106677.

- [13] Sindhubala K, Vijayalakshmi B. Simulation of VLC system under the influence of optical background noise using filtering technique[J]. Materials Today: Proceedings, 2017, 4(2): 4239-4250.
- [14] 李国梁. 基于深度学习的红外与可见光图像融合算法研究[D]. 北京: 北京交通大学, 2021.
- [15] Elaiyaraja K, Kumar M S. Fusion Imaging in Pixel Level Image Processing Technique – A Literature Review[J]. International Journal of Engineering & Technology, 2018,7(3.12):175-177.
- [16] 陈国洋. 基于无监督深度学习的红外与可见光图像融合算法研究[D]. 江苏: 江南大学, 2021.
- [17] Burt P J, Adelson E H. The Laplacian pyramid as a compact image code[M]//Readings in computer vision. Morgan Kaufmann, 1987: 671-679.
- [18] Toet A. Hierarchical image fusion[J]. Machine Vision and Applications, 1990, 3(1): 1-11.
- [19] Burt P J, Kolczynski R J. Enhanced image capture through fusion[C]//1993 (4th) international Conference on Computer Vision. IEEE, 1993: 173-182.
- [20] Liu Z, Tsukada K, Hanasaki K, et al. Image fusion by using steerable pyramid[J]. Pattern Recognition Letters, 2001, 22(9): 929-939.
- [21] Freeman W T, Adelson E H. The design and use of steerable filters[J]. IEEE Transactions on Pattern analysis and machine intelligence, 1991, 13(9): 891-906.
- [22] Chipman L J, Orr T M, Graham L N. Wavelets and image fusion[C]//Proceedings., International Conference on Image Processing. IEEE, 1995, 3: 248-251.
- [23] Zou Y, Liang X, Wang T. Visible and infrared image fusion using the lifting wavelet[J]. TELKOMNIKA Indonesian Journal of Electrical Engineering, 2013, 11(11): 6290-6295.
- [24] Do M N, Vetterli M. The contourlet transform: an efficient directional multiresolution image representation[J]. Official Journal-European Union Information and Notices C, 2006, 49(27A): 2091.
- [25] Yang B, Li S, Sun F. Image fusion using nonsubsampling contourlet transform[C]//Fourth International Conference on Image and Graphics (ICIG 2007). IEEE, 2007: 719-724.

-
- [26] Yang B, Li S. Multifocus image fusion and restoration with sparse representation[J]. IEEE transactions on Instrumentation and Measurement, 2009, 59(4): 884-892.
 - [27] Yao Y, Guo P, Xin X, et al. Image fusion by hierarchical joint sparse representation[J]. Cognitive Computation, 2014, 6: 281-292.
 - [28] Liu Y, Chen X, Ward R K, et al. Image fusion with convolutional sparse representation[J]. IEEE signal processing letters, 2016, 23(12): 1882-1886.
 - [29] Liu Y, Chen X, Peng H, et al. Multi-focus image fusion with a deep convolutional neural network[J]. Information Fusion, 2017, 36: 191-207.
 - [30] Li H, Wu X J, Kittler J. Infrared and visible image fusion using a deep learning framework[C]//2018 24th international conference on pattern recognition (ICPR). IEEE, 2018: 2705-2710.
 - [31] Li H, Wu X, Durrani T S. Infrared and visible image fusion with ResNet and zero-phase component analysis[J]. Infrared Physics & Technology, 2019, 102: 103039.
 - [32] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.
 - [33] Ma J, Yu W, Liang P, et al. FusionGAN: A generative adversarial network for infrared and visible image fusion[J]. Information fusion, 2019, 48: 11-26.
 - [34] Ma J, Xu H, Jiang J, et al. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion[J]. IEEE Transactions on Image Processing, 2020, 29: 4980-4995.
 - [35] Zhang H, Yuan J, Tian X, et al. GAN-FM: Infrared and visible image fusion using GAN with full-scale skip connection and dual Markovian discriminators[J]. IEEE Transactions on Computational Imaging, 2021, 7: 1134-1147.
 - [36] Liu X, Wang R, Huo H, et al. An attention-guided and wavelet-constrained generative adversarial network for infrared and visible image fusion[J]. Infrared Physics & Technology, 2023, 129: 104570.
 - [37] Wang Y, Yao H, Zhao S. Auto-encoder based dimensionality reduction[J]. Neurocomputing, 2016, 184: 232-242.
 - [38] Li H, Wu X J. DenseFuse: A fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2614-2623.

-
- [39] Li H, Wu X J, Durrani T. NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(12): 9645-9656.
 - [40] Li H, Wu X J, Kittler J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images[J]. Information Fusion, 2021, 73: 72-86.
 - [41] Zhao Z, Xu S, Zhang J, et al. Efficient and model-based infrared and visible image fusion via algorithm unrolling[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(3): 1186-1196.
 - [42] Xue W, Wang A, Zhao L. FLFuse-Net: A fast and lightweight infrared and visible image fusion network via feature flow and edge compensation for salient information[J]. Infrared Physics & Technology, 2022, 127: 104383.
 - [43] Tang L, Yuan J, Zhang H, et al. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware[J]. Information Fusion, 2022, 83: 79-92.
 - [44] Tang L, Yuan J, Ma J. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network[J]. Information Fusion, 2022, 82: 28-42.
 - [45] Wang D, Liu J, Liu R, et al. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection[J]. Information Fusion, 2023, 98: 101828.
 - [46] Liu J, Dian R, Li S, et al. SGFusion: A saliency guided deep-learning framework for pixel-level image fusion[J]. Information Fusion, 2023, 91: 205-214.
 - [47] Zhou Z, Fei E, Miao L, et al. A perceptual framework for infrared-visible image fusion based on multiscale structure decomposition and biological vision[J]. Information Fusion, 2023, 93: 174-191.
 - [48] Toet A. TNO image fusion dataset. figshare[J]. 2014
 - [49] Jia X, Zhu C, Li M, et al. LLVIP: A visible-infrared paired dataset for low-light vision[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 3496-3504.
 - [50] Xu H, Ma J, Jiang J, et al. U2Fusion: A unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020,

44(1): 502-518.

- [51] 田鹏飞. 红外热成像气密性检测方法的技术探讨[D]. 扬州: 扬州大学, 2013.
- [52] Al-Wassai F A, Kalyankar N V, Al-Zaky A A. Multisensor images fusion based on feature-level[J]. arXiv preprint arXiv:1108.4098, 2011.
- [53] Du P, Zhang W, Xia J. Hyperspectral remote sensing image classification based on decision level fusion[J]. Chinese Optics Letters, 2011, 9(3): 031002.
- [54] 郑远攀, 李广阳, 李晔. 深度学习在图像识别中的应用研究综述[J]. 计算机工程与应用, 2019, 55(12): 20-36.
- [55] 夏晗. 基于深度学习的红外和可见光图像融合方法研究[D]. 天津: 天津大学, 2020.
- [56] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.
- [57] Egmont-Petersen M, de Ridder D, Handels H. Image processing with neural networks—a review[J]. Pattern recognition, 2002, 35(10): 2279-2301.
- [58] 张绍. 激光主动成像下基于卷积神经网络的目标检测算法研究[D]. 北京: 中国科学院大学(中国科学院长春光学精密机械与物理研究所), 2021.
- [59] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [60] 朱贺. 基于深度学习的图像语义分割算法的应用研究[J]. 电子元器件与信息技术, 2022, 6(02):196-198.
- [61] 周航. 基于深度学习的红外与可见光图像融合[D]. 沈阳: 沈阳理工大学, 2021.
- [62] Ross P J. Taguchi techniques for quality engineering: loss function, orthogonal experiments, parameter and tolerance design[J]. 1988.
- [63] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- [64] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
- [65] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.

-
- [66] Bai X, Zhang Y, Zhou F, et al. Quadtree-based multi-focus image fusion using a weighted focus-measure[J]. *Information Fusion*, 2015, 22: 105-118.
 - [67] Liu Z, Song Y, Sheng V S, et al. MRI and PET image fusion using the nonparametric density model and the theory of variable-weight[J]. *Computer Methods and Programs in Biomedicine*, 2019, 175: 73-82.
 - [68] Yin M, Duan P, Liu W, et al. A novel infrared and visible image fusion algorithm based on shift-invariant dual-tree complex shearlet transform and sparse representation[J]. *Neurocomputing*, 2017, 226: 182-191.
 - [69] Hu H M, Wu J, Li B, et al. An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels[J]. *IEEE Transactions on Multimedia*, 2017, 19(12): 2706-2719.
 - [70] Panigrahy C, Seal A, Mahato N K. Fractal dimension based parameter adaptive dual channel PCNN for multi-focus image fusion[J]. *Optics and Lasers in Engineering*, 2020, 133: 106141.
 - [71] Wang Y, Wang B S, Wang T, et al. Image entropy active contour models towards water area segmentation in remote sensing image[J]. *Optics and Precision Engineering*, 2018, 26(3): 698-707.
 - [72] Xydeas C S, Petrovic V. Objective image fusion performance measure[J]. *Electronics letters*, 2000, 36(4): 308-309.
 - [73] Aslantas V, Bendes E. A new image quality metric for image fusion: The sum of the correlations of differences[J]. *Aeu-international Journal of electronics and communications*, 2015, 69(12): 1890-1896.
 - [74] Ma J, Zhou Z, Wang B, et al. Infrared and visible image fusion based on visual saliency map and weighted least square optimization[J]. *Infrared Physics & Technology*, 2017, 82: 8-17.
 - [75] Chen Y, Fan H, Xu B, et al. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution[C]//*Proceedings of the IEEE/CVF international conference on computer vision*. 2019: 3435-3444.
 - [76] Li Y, Yao T, Pan Y, et al. Contextual transformer networks for visual recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 45(2): 1489-1500.

- [77] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.
- [78] Gao C, Song C, Zhang Y, et al. Improving the performance of infrared and visible image fusion based on latent low-rank representation nested with rolling guided image filtering[J]. IEEE Access, 2021, 9: 91462-91475.
- [79] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Proceedings of the AAAI conference on artificial intelligence. 2017, 31(1).
- [80] Hwang S, Park J, Kim N, et al. Multispectral pedestrian detection: Benchmark dataset and baseline[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1037-1045.
- [81] Toet A. Image fusion by a ratio of low-pass pyramid[J]. Pattern recognition letters, 1989, 9(4): 245-253.
- [82] Ma J, Zhang H, Shao Z, et al. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-14.
- [83] Zhang R, Zheng K, Shi P, et al. Traffic sign detection based on the improved YOLOv5[J]. Applied Sciences, 2023, 13(17): 9748.

攻读硕士学位期间发表学术论文及参研项目

一、发表学术论文

- [1] Shi P, **Mao F**, Zhang R. DAE-Nest: A depth information extraction and enhancement fusion network for infrared and visible images[J]. Optics Communications, 2024, 560: 130441. (中科院三区, 除导师外一作, 见刊)
- [2] Shi P, **Mao F**, Zhang R. Infrared and Visible Images Fusion Method Based on Dual-Scales and Attention Mechanism[J]. Journal of Intelligent & Fuzzy Systems, 2024 (中科院四区, 除导师外一作, 终审)

二、专利

- [1] 一种基于新能源汽车电池的充电装置及其方法[P]. 中国: CN114905993A, 2022-09-02. (发明, 除导师外一作, 实审)

三、参研项目

- [1] 安徽省自然科学基金 (2208085MF173);
- [2] 安徽省重点研发计划项目 (202104a05020003);
- [3] 长三角科技创新共同体联合攻关项目(2023CSJGG1600)
- [4] 基于自主学习的无人驾驶决策方法研究, 安徽工程大学研究生实践与创新基金项目

致谢

山高路远，行则将至。转眼之间十余年求学生涯已至尾声，当我站在这段旅程的终点，回望过去那些给予我帮助的人，心中充满了无限的感激之情。

首先，我要感谢我的研究生导师时培成教授，时教授不仅是我的学术引导者，更是我们课题组的明灯。在我做研究的每一个难关，每一个疑惑中，他总是耐心指导，用独有的智慧和经验引领我前行。时教授深深的教诲，不仅让我在学术上有所发现，更让我在成长的道路上受益匪浅。

其次，我要感谢我的家人，在我求学的十余年间，给予了我无数的温暖与支持。在这段学习和研究的过程中，是她们给予了我无限的支持和鼓励。在我遇到困难和挑战时，是她们的理解和爱抚平了我的焦虑，给我力量继续前行。她们的支持是我不断前进的动力源泉。

然后，我还要感谢我的同门张建国、江彤、张庆、李屹，感谢齐恒师兄、刘志强师兄、陈新禾师兄、张程辉师兄、刘糠继师兄、李龙师兄以及师弟师妹们，是他们的陪伴和帮助让我的研究之路不再孤单。课题组之间共同讨论，互相启发，共同进步。这段旅程因为有他们而变得更加丰富和有意义。

最后，我要感谢所有直接或间接帮助和支持我的人。每一次的交流和互动都是我宝贵的财富，让我在这段学术旅程中不断成长和进步。时间如逆旅，我亦如行人。感谢所有人的陪伴和支持，让我能够勇敢地走完这段旅程。未来，我将带着这份感激继续前行，探索更广阔的天地。

尚德敏学 唯实惟新

