

中图法分类号: TP37 文献标识码: A 文章编号: 1006-8961(2023)01-0102-16

论文引用格式: Zuo Y F, Fang Y M and Ma K D. 2023. The critical review of the growth of deep learning-based image fusion techniques. Journal of Image and Graphics, 28(01): 0102-0117 (左一帆, 方玉明, 马柯德. 2023. 深度学习时代图像融合技术进展. 中国图象图形学报, 28(01): 0102-0117) [DOI:10.11834/jig.220556]

深度学习时代图像融合技术进展

左一帆¹, 方玉明^{1*}, 马柯德²

1. 江西财经大学信息管理学院, 南昌 330013; 2. 香港城市大学计算机学院, 香港 999077

摘要: 为提升真实场景视觉信号的采集质量, 往往需要通过多种融合方式获取相应的图像, 例如, 多聚焦、多曝光、多光谱和多模态等。针对视觉信号采集的以上特性, 图像融合技术旨在利用同一场景不同视觉信号的优势, 生成单图像信息描述, 提升视觉低、中、高级任务的性能。目前, 依托端到端学习强大的特征提取、表征及重构能力, 深度学习已成为图像融合研究的主流技术。与传统图像融合技术相比, 基于深度学习的图像融合模型性能显著提高。随着深度学习研究的深入, 一些新颖的理论和方法也促进了图像融合技术的发展, 如生成对抗网络、注意力机制、视觉 Transformer 和感知损失函数等。为厘清基于深度学习技术的图像融合研究进展, 本文首先介绍了图像融合问题建模, 并从传统方法视角逐渐向深度学习视角过渡。具体地, 从数据集生成、神经网络构造、损失函数设计、模型优化和性能评估等方面总结了基于深度学习的图像融合研究现状。此外, 还讨论了选择性图像融合这类衍生问题建模(如基于高分辨率纹理图融合的深度图增强), 回顾了一些基于图像融合实现其他视觉任务的代表性工作。最后, 根据现有技术的缺陷, 提出目前图像融合技术存在的挑战, 并对未来发展趋势给出了展望。

关键词: 图像融合; 深度神经网络; 随机梯度优化(SGD); 图像质量评价; 深度学习

The critical review of the growth of deep learning-based image fusion techniques

Zuo Yifan¹, Fang Yuming^{1*}, Ma Kede²

1. School of Information Management, Jiangxi University of Finance and Economics, Nanchang 330013, China;

2. Department of Computer Science, City University of Hong Kong, Hong Kong 999077, China

Abstract: To capture more effective visual information of the natural scenes, multi-sensor imaging systems have been challenging in multiple configurations or modalities due to the hardware design constraints. It is required to fuse multiple source images into a single high-quality image in terms of rich and feasible perceptual information and few artifacts. To facilitate various image processing and computer vision tasks, image fusion technique can be used to generate a single and clarified image features. Traditional image fusion models are often constructed in accordance with label-manual features or unidentified feature-learned representations. The generalization ability of the models needs to be developed further. Deep learning technique is focused on progressive multi-layer features extraction via end-to-end model training. Most of demonstration-relevant can be learned for specific task automatically. Compared with the traditional methods, deep learning-based models can improve the fusion performance intensively in terms of image fusion. Current image fusion-related deep learning models

收稿日期: 2022-05-31; 修回日期: 2022-10-09; 预印本日期: 2022-10-16

* 通信作者: 方玉明 leo.fangyuming@foxmail.com

基金项目: 国家自然科学基金项目(62071407, 62132006, 61901197); 深圳市科技创新基金项目(2021S2vup051)

Supported by: National Natural Science Foundation of China (62071407, 62132006, 61901197); Science Innovation Foundation of Shenzhen (2021S2vup051)

are often beneficial based on convolutional neural networks (CNNs) and generative adversarial networks (GANs). In recent years, the newly network structures and training techniques have been incorporated for the growth of image fusion like vision transformers and meta-learning techniques. Most of image fusion-relevant literatures are analyzed from specific multi-fusion issues like exposure, focus, spectrum image, and modality issues. However, more deep learning-related model designs and training techniques is required to be incorporated between multi-fusion tasks. To draw a clear picture of deep learning-based image fusion techniques, we try to review the latest image fusion researches in terms of 1) dataset generation, 2) neural network construction, 3) loss function design, 4) model optimization, and 5) performance evaluation. For dataset generation, we emphasize two categories: a) supervised learning and b) unsupervised (or self-supervised) learning. For neural network construction, we distinguish the early or late stages of this construction process, and the issue of information fusion is implemented between multi-scale, coarse-to-fine and the adversarial networks-incorporated (i.e., discriminative networks) as well. For loss function design, the perceptual loss functions-specific method is essential for image fusion-related perceptual applications like multi-exposure and multi-focus image fusion. For model optimization, the generic first-order optimization techniques are covered (e.g., stochastic gradient descent (SGD), SGD + momentum, Adam, and AdamW) and the advanced alternation and bi-level optimization methods are both taken into consideration. For performance evaluation, a commonly-used quantitative metrics are reviewed for the manifested measurement of fusion performance. The relationship between the loss functions (also as a form of evaluation metrics) are used to drive the learning of CNN-based image fusion methods and the evaluation metrics. In addition, to illustrate the transfer feasibility of image fusion-consensus to a tailored image fusion application, a selection of image fusion methods is discussed (e.g., a high-quality texture image-fused depth map enhancement). Some popular computer vision tasks are involved in (such as image denoising, blind image deblurring, and image super-resolution), which can be resolved by image fusion innovatively. Finally, we review some potential challenging issues, including: 1) reliable and efficient ground-truth training data-constructed (i.e., the input image sequence and the predictable image-fused), 2) lightweight, interpretable, and generalizable CNN-based image fusion methods, 3) human or machine-related vision-perceptual calibrated loss functions, 4) convergence-accelerated image fusion models in related to adversarial training setting-specific and the bias-related of the test-time training, and 5) human-related ethical issues in relevant to fairness and unbiased performance evaluation.

Key words: image fusion; deep neural networks; stochastic gradient descent (SGD); image quality assessment; deep learning

0 引言

由于硬件和成像机理限制,视频图像采集系统往往需要使用不同设备和配置对真实场景进行多次采集,以捕捉特定视觉场景更多更全的细节。针对以上方式获取的不同图像描述,图像融合旨在综合各个图像中的优势信息表征,形成高质量的单图像描述,为各类计算机视觉任务提供助力,例如计算摄影(Petschnigg等,2004)、图像复原(Deng和Dragotti,2021)、图像分割(Minaee等,2022)、目标跟踪(Zhu等,2019)、医学诊断(Bhatnagar等,2013)和遥感监测(Ghassemian,2016)等。

图像融合技术研究属于底层计算机视觉任务,其研究起步较早,但始终是热门研究领域。传统图像融合的主要应用场景包括以下4个方面。1)多聚焦图像融合。将同一场景聚焦区域不同的多幅图

像融合成一幅全清晰图像。2)多曝光图像融合。将不同曝光图像的有效信息结合起来,拓展曝光成像的细节及动态范围。3)高光谱图像融合。将低空间分辨率高光谱图像与同一场景高空间分辨率的多光谱图像结合,提高高光谱图像空间分辨率。4)多模态图像融合。将不同模态(医学)图像互补信息融合,实现对成像场景更全面的描述。

深度学习出现以前,传统图像融合技术的研究主要经历了两个时期,即基于数学理论的预定义模型和基于数据驱动的浅层学习模型。具体来说,由于数据匮乏,图像融合技术的研究首先基于严谨的数学理论构造模型,并在空域和变换域调整融合程度和策略。该时期的主流技术包括:边缘滤波(Li等,2013;Liu等,2021b)、层次分解(Son和Zhang,2016)、多尺度变换(Wang等,2014;Yang等,2010)、成分编辑(Vivone,2019;Yokoya等,2012)和变分优化(Meng等,2019;Ballester等,2006)等。然

而,由于以上模型的构建依据特定假设,故当假设与实际场景不符时,模型的泛化能力明显受限,导致其在实际场景应用中性能不尽人意。一个典型的问题是预定义模型对不同来源的图像使用相同的函数提取和融合特征,未综合考虑各类图像的数据分布特点。随着采集数据的逐步丰富,图像融合技术的研究逐渐由模型驱动方式转向数据驱动方式,其主流机器学习模型为:稀疏编码(Zhu 和 Bamler, 2013; Wei 等, 2015; 王丽芳等, 2019)、低秩先验(Li 等, 2020a; Wang 等, 2021)等。虽然引入数据驱动思想有效提升了模型表征能力,然而,以互斥特征学习为主旨的浅层特征表示无法高效地拟合真实场景中复杂的数据分布。

与上述传统方法不同,深度学习通过端对端模型构建,旨在实现特征的逐步提取、抽象,从而学习到与当前任务最相关的表征。早期的研究往往将传统模型的部分结构神经网络化,并根据设计的损失函数自适应调整特征融合方式(Li 等, 2020b; Jin 等, 2022)。随着研究的深入,现代深度学习模型广泛采用端对端的学习范式,可同时考虑特征提取和融合,实现各步骤的流水线协同优化。相比传统方法,基于深度学习的图像融合模型显著提升了融合性能(Yang 等, 2017; Xu 等, 2021)。

目前,在图像融合研究领域广泛使用的深度学习模型包括卷积神经网络(convolutional neural net-

works, CNN)、生成对抗网络(generative adversarial network, GAN)等。其中,作为计算机视觉基础框架的 CNN,主要通过对分级式卷积核的学习以实现特征提取和图像重建(Xu 等, 2020; Xie 等, 2022)。而基于 GAN 的网络框架使用鉴别器学习特定域的数据分布流形,以提升融合图像的主观质量(Guo 等, 2019; Ma 等, 2020a)。另外,由标准 GAN 模型演化而来的多鉴别器结构对融合每个域均使用一个鉴别器,以促使生成器学习多个域的数据分布流形(Xie 等, 2021; Liu 等, 2021a)。同时,深度学习时代的另一个特点是通用框架的构建。传统方法往往在图像融合各子问题领域都有独特的模型设计,然而,深度神经网络的出现使相同的模型框架可用于多个图像融合子任务。该类通用框架的训练可通过设置特殊的训练集对每个任务独立训练(Dudhane 等, 2022),也可在单模型训练中统一考虑多个相关任务,即多任务学习(Li 等, 2021)。

随着深度学习理论研究的不断深入,一些新颖的网络结构和训练技巧促进了图像融合技术的发展,例如,视觉 Transformer 结构。视觉 Transformer 受自然语言处理领域启发,该新型结构是对 CNN 模型的一种补充,其弱化了模型偏置,并强化了远距离依赖建模能力(Su 等, 2022; Qu 等, 2021)。相关基于深度神经网络的一些代表性工作出现的时间线如图 1 所示。

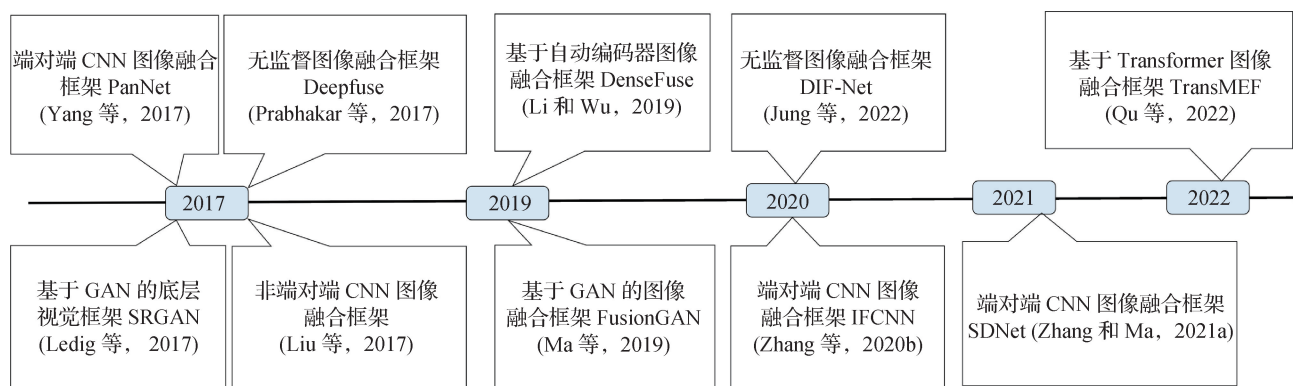


图 1 基于深度神经网络的代表性工作

Fig. 1 Representative image fusion techniques based on deep neural networks

以上图像融合的应用领域共享相同的问题建模,即参与融合的各图像分量地位对等,融合的目标为综合所有信息生成更细致的图像表示。事实上,还有一些特殊任务可看成选择性图像融合问题,即

参与融合的各图像间具有不同的物理背景含义。如果不加限制融合,会导致结果出现歧义,与场景真实情况不符。在这类问题中,还有一类变体,即待融合图像之间存在明确的目标域和指导域标签,其通过

选择性融合指导域信息增强目标域图像质量,例如,基于高分辨率纹理图融合的低分辨率深度图增强等(Ye等,2020;Kim等,2021)。此外,一些低级和高级视觉任务也可以建模为图像融合问题(Zheng等,2020;Tang等,2022a;Xiao等,2022)。

目前,已有的综述类文献未全面从模型设计和训练技巧层面给出基于深度学习的图像融合研究现状。同时,由于深度学习技术迭代更新快,有部分新技术未得到深入分析。为厘清基于深度学习的图像融合技术的最新进展,本文从数据集生成、神经网络构造、损失函数设计、模型优化和性能评估等方面分析图像融合任务的最新进展。此外,不同于上述图像融合问题,本文还讨论了选择性图像融合及其变体定向选择性图像融合等衍生问题建模。另外,本文回顾了一些基于图像融合实现其他视觉任务的代表性工作。

1 基于深度学习的图像融合进展

从数据集生成、神经网络构造、损失函数设计、模型优化和性能评估等方面分析基于深度学习的图像融合领域最新进展。

1.1 数据集生成

根据监督学习和非监督学习的类别,分别阐述数据集生成技巧。

1) 监督学习。对于监督学习范式,训练集的组成显式包含 Ground-truth。然而,绝大多数图像融合任务存在 Ground-truth 难获取的问题。因此,在监督学习的范式下,如何高效准确地获取大量 Ground-truth 融合图像是任务成功的关键。以多曝光图像融合为例,Cai 等人(2018)通过对 13 种多曝光融合算法生成的图像进行主观实验,挑选出当前场景感知质量最好的融合图像作为 Ground-truth。该过程因需要依赖人类标注难以规模化。以多聚焦、多光谱图像融合为例,Ground-truth 图像的获取往往依赖昂贵精密的硬件设备,如光场相机(light-field cameras),机载传感器(airborne sensors)等,这极大地限制了训练集的规模。再以多模态医学图像融合为例,Ground-truth 图像甚至没有清晰明确的定义。不少文献(Hermessi等,2021)以对比度作为 Ground-truth 图像选取标准。然而,对比度与计算机辅助诊断(computer-aided diagnosis)的精度并非线性相关。

鉴于此,无监督学习范式逐渐受到图像融合领域学者重视。

2) 无监督学习。无监督学习范式的成败则极大地取决于输入图像序列的质量。当然,不同图像融合应用对“质量”的定义有所不同。以多曝光图像融合为例,仅依赖于极端欠曝光和过曝光两幅图像作为输入算法(Li和Wu,2019)难以在富有挑战性的高动态范围场景取得好的结果。Gupta等人(2013)提出基于斐波那契数列的多重曝光值选择,以期在融合算法固定的情况下达到最佳融合效果。Wang等人(2020)则引入强化学习的思想进行多重曝光值选择。以多聚焦图像融合为例,输入图像序列大多来源于对一幅清晰图像的不同局部区域进行不同程度的高斯模糊。众所周知,失焦模糊与景深密切相关。因此,基于光场图像(Jin等,2020)和合成失焦技术(Barron等,2015)的多聚焦图像序列生成值得重新关注。以光谱图像融合为例,研究者一般根据空域和光谱域观察模型对高分辨率高光谱图像下采样,得到低分辨率高光谱图像和高分辨率多光谱图像(或对应的全色图像),作为网络模型的输入。而多模态图像融合任务的输入由于受到硬件采集设备的限制,需选取构建具有互补信息的图像序列,如红外图像(用于捕捉低照度下的发热源)和可见光图像(用于捕捉物体颜色和结构)融合(Ma等,2019;刘明葳等,2021);CT(computed tomography)图像(用于描述硬组织结构)、MRI(magnetic resonance imaging)图像(用于描述软组织结构)、PET(positron emission tomography)图像(用于描述脏器功能和新陈代谢)和SPECT(single-photon emission computed tomography)图像(用于描述脏器血液循环情况)融合(Zhang等,2020b)。最后,值得指出的是,图像融合往往依赖于图像配准。在大多数图像融合任务的源图像序列生成过程中,图像欠配准问题一般不大,可用经典的基于空域或特征域的全局匹配算法进行粗到精的配准。RFNet(Xu等,2022b)则提出了配准与融合联合优化的方法,是将配准任务直接服务于融合任务。

1.2 神经网络构造

神经网络构造分为非端对端模式和端对端模式。

1) 非端对端神经网络。一般模型将图像融合建模成源图像间的加权和。早期基于深度神经网络

的图像融合方法仅使用 CNN 监督学习该加权系数图,而非直接对最终融合图像进行约束。一个典型的例子是将多聚焦图像融合建模成二分类任务(Liu 等,2017)。因此,该类模型采用非端对端训练方式,典型的网络框架如图 2 所示(Li 等,2020b;Liu 等,2017)。

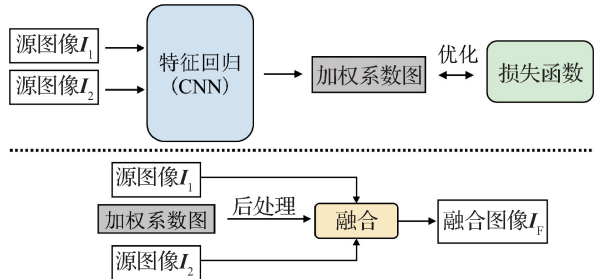


图 2 图像融合任务的非端到端方式 CNN 网络

Fig. 2 Non-end-to-end CNN for image fusion

为促进加权系数图更贴近真实情况,如增加加权系数图边缘与源图像边缘耦合性等,Guo 等人(2019)引入了鉴别器用于区分生成器伪造的加权系数图和高质量的 Ground-truth 系数图。该类方法的通用框架如图 3 所示。

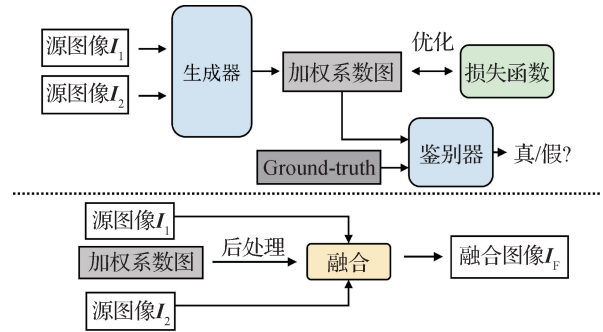


图 3 图像融合任务的非端到端方式 GAN 网络

Fig. 3 Non-end-to-end GAN for image fusion

虽然,神经网络工具的引入极大地促进了图像融合研究的进步,但是,受传统模型的限制,非端到端训练的神经网络无法对特征提取与融合各阶段协同优化,限制了性能的持续提升。再者,Ground-truth 系数图的构造没有统一标准,缺乏基于人类视觉或机器视觉的目标函数指引。

2)端对端神经网络。为克服非端到端模型的弊端,国内外研究者提出了大量的端对端神经网络。该类模型大致分为 3 类,相应的框架如图 4 所示。第 1 类网络仅将输入的多幅源图像在通道维度上合

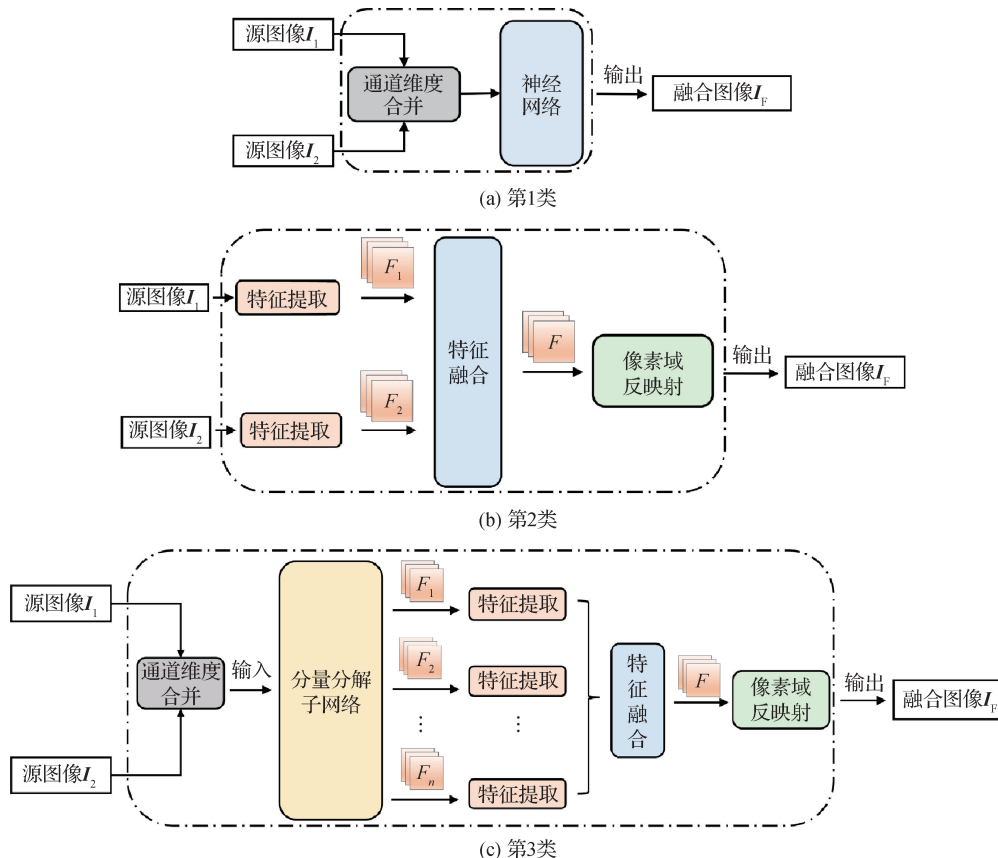


图 4 图像融合任务的主流端到端 CNN 网络结构

Fig. 4 Mainstream end-to-end CNN architectures for image fusion((a) the first class;(b) the second class;(c) the third class)

并后直接送入神经网络中(Xu等,2022a;Xu和Ma,2021),属于源图像像素域早期融合。第2类网络采用双头分支,首先分别对所有源图像进行特征提取,然后在学到的特征空间实现后期融合,最后将融合的特征经过重构模块映射至图像像素域空间。通常而言,这种先分后总的设计有利于神经网络学习到高阶特征,提升其特征融合效率,进而扩展模型的泛化性。根据源图像是否跨域,上述神经网络的特征提取分别采用参数共享(Jung等,2020;Li和Wu,2019;Zhang等,2020b;Xie等,2022)和非参数共享(Fu等,2019;Qu等,2018;Zheng等,2021;Zhao等,2021;Jin等,2022)方式,以优化特征提取效果并兼顾计算复杂度。此外,该类框架还存在一类变体,即在特征提取阶段获得的各源图像特征之间引入通信机制。基于已有的特征表示,该设计有利于增强各源图像特征的差异性表达,提升特征提取的质量(Tang等,2022b;Yao等,

2020;Zhang等,2020b)。第3类网络采用总分总结构,即在第2类网络框架的基础上新增分量分解子网络。该步骤将通道维度上合并的源图像作为分量分解子网络的输入,例如,低频高频分解(Yang等,2017;Hu等,2021)、空域光谱分解(Xu等,2021)等。然后采用第2类框架针对各类分量特点设计特征提取模块,并衔接特征融合和像素域反映射实现图像融合。

在上述3类框架上,研究者还显式地在特征提取阶段和特征融合阶段分别引入或同时引入多尺度、渐进式的特征融合策略,以提升特征质量(Zhao等,2019;Xu等,2020;Zhang等,2021b),对应的框架示意图如图5所示。图3—图5中特征提取、融合网络用到的核心技术包括可导联合引导滤波(Ma等,2020c)、残差连接(Xu等,2022a)、参数共享(Jung等,2020)、多尺度结构(Hu等,2021)和低秩分解(Xie等,2022)等。

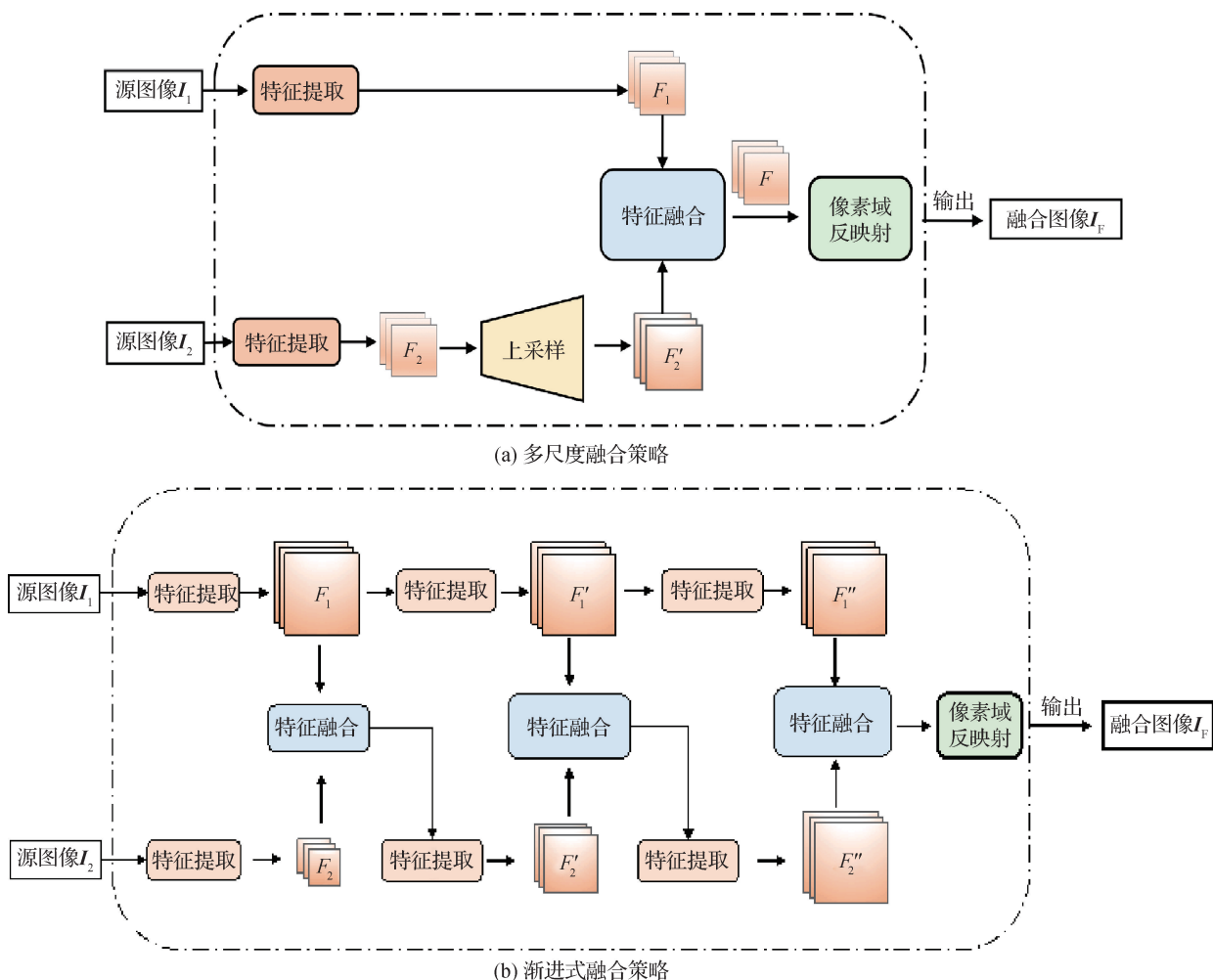


图5 多尺度和渐进式的融合策略

Fig.5 Multi-scale and progressive feature fusion strategy

((a) muti-scale feature fusion strategy; (b) progressive feature fusion strategy)

最近,由于受到自然语言处理领域的影响,除了以上基于 CNN 的框架,研究者还提出了一些基于 Transformer 的图像融合框架 (Qu 等, 2021; Su 等, 2022)。特别地,基于当前主流 Swin Transformer, SwinFusion 提出一种通用的图像融合方法,可以有效解决多模态、多聚焦和多曝光等各种图像融合任务 (Ma 等, 2022)。

目前,Transformer 能表现出比 CNN 框架更好效果的原因尚不明确,现在普遍被接受的分析主要体现在:1) Transformer 具有强大的远距离依赖建模能力,即感受野比传统 CNN 要大。特别地,全局 Transformer (Dosovitskiy 等, 2021) 的感受野是整个图像,局部 Transformer (Liu 等, 2021b) 中的窗口尺寸也大于 CNN 常用的 3×3 卷积核。虽然,随着大量 3×3 卷积核的堆叠可在理论上增加感受野,但实

际感受野距离理论值具有显著的差距 (Ding 等, 2022)。2) Transformer 的计算方式与低通滤波器相似,可实现模型集成学习机制。相对地, CNN 的工作机理更偏向于高通滤波器。从这个角度说,综合 CNN 和 Transformer 的优势是一个值得探索的新方向 (Park 和 Kim, 2022)。

与非端对端神经网络类似,端对端神经网络中也可引入鉴别器,形成生成对抗网络,其目标为引导融合图像数据分布向真实图像数据分布流形靠近。相应方法的通用框架如图 6 所示,其设计可分为传统的单鉴别器 (Ma 等, 2019; Liu 等, 2021a; Xie 等, 2021) 和多鉴别器 (Ma 等, 2020a, b) 两类。其中,多鉴别器可使融合图像最大限度地保留各源图像信息,其典型应用为可见光图像与红外图像的融合、高光谱图像融合和医学图像融合等。

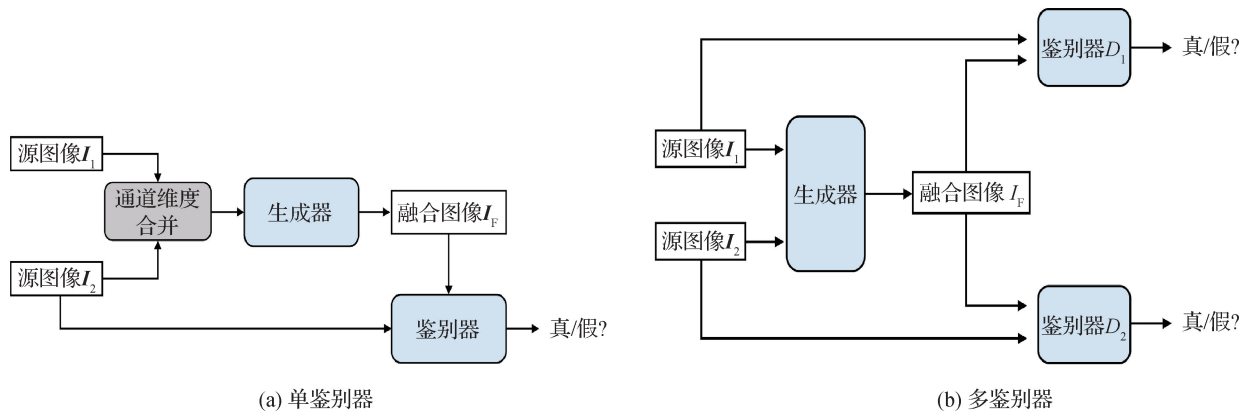


图 6 图像融合任务的端到端生成对抗网络

Fig. 6 End-to-end GANs for image fusion((a) single discriminator; (b) dual discriminators)

1.3 基于感知的损失函数设计

本节讨论面向图像融合的损失函数设计,其分为通用部分和问题相关的特殊部分。特别说明,为防止过拟合的参数正则化等技巧不在本节讨论范围。一般来说,图像融合领域的损失函数往往包含多项,各项之间采用加权和方式结合。

1) 通用损失函数。一些常规的损失函数使用 L_1 范数、 L_2 范数和结构相似性 (structural similarity, SSIM) (Wang 等, 2004) 等评价融合图像 I_f 与参考图像 I_g 在像素域的相似度 (Xu 和 Ma, 2021; Fu 等, 2019; Tang 等, 2022a; Xu 等, 2022b; Li 和 Wu, 2019; Yang 等, 2017)。其中, L_2 范数家族包括峰值信噪比 (peak signal-to-noise ratio, PSNR)、均方根误差 (root mean square error, RMSE) 和均方误差 (mean square

error, MSE) 等价准则。参考图像在监督学习中是 Ground-truth, 在无监督学习中设计灵活,例如,自监督中输入的红外图像和可见光图像等。值得注意的是,在非端对端网络架构下,监督信息需针对加权系数图设计损失函数 (Liu 等, 2017)。

随着研究的深入,研究者发现单独使用上述基础损失函数易导致过度模糊的融合图像。为提升融合图像的主观质量,研究者还尝试在损失函数中附加对梯度域 (Li 等, 2020b, 2021; Zhang 等, 2020a; Xu 等, 2022b)、深层特征域 (Zhang 等, 2020b) 的评价,其对应的相似度评价函数一般使用 L_1 范数、 L_2 范数构建。其中,深层特征使用预训练模型前向传播获取特定层的输出特征,例如 VGG (Visual Geometry Group) 网络 (Simonyan 和 Zisserman, 2015)。此外,

通过联合监督网络输出和中间特征状态, 损失函数可基于上述通用设计对多尺度、多阶段的特征深度监督(Su 等, 2022; Zhao 等, 2019)。

进一步地, 通过引入鉴别器 D , 上述生成网络 G 可扩展为生成对抗网络。其中, 鉴别器的常规设计为分类器, 以区分生成器输出的伪造图像和 Ground-truth, 即促进生成器学习真实图像数据分布流形。从这个角度出发, 整个对抗网络中需要新增一个分类器损失函数 L_{GAN} , 常用设计为交叉熵函数, 如式(1)所示(Ma 等, 2020a; Liu 等, 2021a; Guo 等, 2019)。

$$L_{\text{GAN}} = \min_{G_\theta} \max_{D_\theta} E_{x \sim p_{\text{real}}} \log D(x) + E_{y \sim p_{\text{real}}} \log(1 - D(G(y))) \quad (1)$$

式中, G_θ 和 D_θ 分别为生成器 G 和鉴别器 D 的可学习参数集合, p_{real} 和 p_{fake} 是真实图像数据分布和待融合图像输入的数据分布。

2) 先验损失函数。除了以上生成网络通用损失函数设计外, 一些工作显式地考虑了传统图像先验, 以辅助网络学习真实图像数据分布流形, 例如, 全变差(total variation, TV)范数损失函数(Qu 等, 2021)、结构张量损失函数(Jung 等, 2020)。该结构张量先验属于梯度域先验变种, 具体定义参看 Di Zenzo(1986)文献。

另外, 在光谱图像融合任务中, 高光谱图像和多光谱图像之间的丰度稀疏性先验 L_{abs} 和基于皮尔逊相关系数的光谱相似度度量 L_{spec} 都可作为损失函数项, 其定义如式(2)(3)所示(Zheng 等, 2021; Yao 等, 2020; Hu 等, 2021; Xu 等, 2021)。

$$L_{\text{abs}} = \sum_{i,j} KL(a \| \tilde{a}_{i,j}) \quad (2)$$

式中, KL 表示 KL 散度(Kullback-Leibler divergence)度量, a 是一个小常数, $\tilde{a}_{i,j}$ 为丰度矩阵元素。

$$L_{\text{spec}} = \frac{1}{hw} \sum_{i=0}^h \sum_{j=0}^w \left| \frac{\langle I_f(i,j), I_g(i,j) \rangle}{\|I_f(i,j)\|_2 \|I_g(i,j)\|_2} - 1 \right| \quad (3)$$

式中, h 和 w 是多光谱图像的高度和宽度。 I_f 和 I_g 分别代表融合图像和参考图像。

在生成对抗网络损失函数设计中, 研究者尝试了其他变种, 例如, 最小二乘生成对抗损失 L_{ISGAN} (Xie 等, 2021; Ma 等, 2020b, 2019)。最小二乘生成对抗为

$$L_{\text{ISGAN}} = \min_{G_\theta} \min_{D_\theta} \frac{1}{2} E_{x \sim p_{\text{real}}} (D(x) - l_{\text{real}})^2 +$$

$$\frac{1}{2} E_{y \sim p_{\text{fake}}} (D(G(y)) - l_{\text{fake}})^2 \quad (4)$$

式中, l_{real} 和 l_{fake} 分别代表真实样本和生成样本的标签。

上述 L_{ISGAN} 有利于推动生成器生成的样本落在分类决策面附近, 且能提供稳定的梯度, 提升训练鲁棒性。

3) 基于感知的损失函数。对于多曝光、多聚焦等图像融合任务, 生成图像的质量好坏最终由人眼判定。因此, 驱动神经网络学习的损失函数需尽可能满足人眼感知特性, 即计算的函数值与人类主观偏好一致。目前, 该方向研究乏善可陈。Ma 等人(2018, 2020c)将 SSIM 拓展到多曝光图像融合领域, 并成功将其用于多曝光图像融合的感知优化。类似地, Rahman 等人(2017)将视觉信息保真度(visual information fidelity, VIF)指标扩展到多曝光图像融合领域, 提高了融合图像感知质量预测性能。上述两种指标对图像细节、对比度均有所度量, 因此也有希望在多聚焦图像融合任务的感知优化上有所建树。对于多模态、高光谱图像融合任务, 生成图像的质量好坏往往由下游机器学习任务判定。因此, 驱动神经网络学习的损失函数需满足机器感知特性, 亦即下游机器学习任务的精度。目前, SeAFusion(Tang 等, 2022a)和 TarDAL(Liu 等, 2022)是以提升下游任务精度为目标的图像融合领域的开创性工作。SeAFusion 选择动态调整底层任务和下游任务损失函数间的加权系数。而 TarDAL 则通过引入两级优化设计了以检测为目标的红外—可见光图像融合模型, 并公开了跨域图像目标检测的首个数据集。

1.4 模型训练

针对通用模型训练方法和训练方法进展两方面展开阐述。

1) 通用训练方法。单生成网络训练的通用方法由优化器选择和学习率调整两个方面组成。广泛采用的优化器包括 SGD(stochastic gradient descent)、SGD + momentum、Adam、RMSProp, 学习率调整包括均匀减半(Jung 等, 2020)、线性衰减(Zheng 等, 2021)、指数衰减(Zhang 等, 2020b)和余弦退火(Loshchilov 和 Hutter, 2017)。值得一提的是, Adam 优化器还存在一个变种 AdamW, 其目的为提升 Adam 对具有 weight decay 正则的损失函数训练稳

定性。此外,对于较难优化或计算开销较大的部分损失函数项,一些工作采用多阶段课程学习训练方式,即先预训练容易优化的损失项,然后对模型微调以联合训练所有项(Zhang等,2020b;Li等,2021)。一些工作提出的复合网络包含多个单网络,其主流训练方式为序列化依次进行(Tang等,2022b;Xu等,2021;Yang等,2017)。

针对由生成器和鉴别器组成的生成对抗网络包括多个网络优化的特点,相关工作一般采用交替训练每个网络的方式。其中,鉴别器往往优先被训练更新,且对生成器和鉴别器的训练次数不一定均衡,训练次数比依据具体问题而异(Ma等,2019;Liu等,2021a;Xie等,2021;Guo等,2019)。

一些工作采用了多任务学习框架,其主流工作采用联合训练和序列训练。联合训练是把所有任务的训练数据糅合在一起,每次迭代随机选取所有任务的部分训练样本。随着任务数量增加,此种方法对运算资源和存储资源都提出了较大挑战。序列训练是按一定顺序依次训练每个任务。该设置虽然可降低资源开销,但随着训练的进行,模型会遗忘相对靠前任务学习的参数,导致训练的模型在近期训练任务上表现良好,而在早期训练任务上性能大幅度下降。

2)训练方法进展。针对互补多任务学习框架序列训练的难点,一些工作创新性地利用元学习,先联合训练多任务模型,而后冻住其他任务支路。针对主任务调整损失函数,对该任务支路模型做适当微调(Li等,2021;Xu等,2022a)。此外,一些工作提出伴随不同任务中模型的训练,在对应已学习的所有模型参数之间引入加权平均。其中,参数权重体现所有参数在对应任务模型中的重要程度,即重要的参数抑制“遗忘”,相对不重要的参数可赋予更大的灵活度,以适配新的任务特点(Xu等,2022a)。

训练方法另一个方面进展体现在生成对抗网络的训练。一些代表性工作简述如下。与传统固定生成器和鉴别器训练次数比不同,该比例可自适应设置(Ma等,2020a),即当鉴别器已区分不出生成器的伪造数据时,可提高鉴别器的训练次数,反之亦然。此外,对于包含多鉴别器的特殊生成对抗网络,一些工作将训练问题拆分成一系列标准生成对抗网络训练子问题,即每次子问题均交替训练公共的生成器和一个特定的鉴别器(Ma等,2020b)。

1.5 评价指标

从图像融合多个应用出发,总结其中的全参考评价指标和无参考评价指标。在下述指标公式中,统一使用 I_f 和 I_g 分别表示融合图像和参考图像, μ 和 σ 代表图像均值和标准差。参考图像可为真实Ground-truth,也可以是输入图像。

1)全参考评价指标。当参考图像可用时,融合图像可通过与对应的参考图像进行相似度比较实现质量评价。相关的评价指标为RMSE、PSNR、SSIM、线性相关系数(correlation coefficient, CC)、互信息(mutual information, MI)、边界相似度(edge-dependent fusion quality index, QE)等。部分公式定义如下

$$MI = \sum_{I_f, I_g} h_{I_f I_g} \log \frac{h_{I_f I_g}}{h_{I_f} h_{I_g}} \quad (5)$$

式中, $h_{I_f I_g}$ 为参考图像和融合图像联合直方图,类似地, h_{I_f} 和 h_{I_g} 分别是融合图像和参考图像边缘直方图。该评价指标一定程度上减弱了图像整体亮度差异、配准误差对相似度评估的影响,侧重考虑线性映射关系假设下的相似度。

$$QE = \frac{\sum_{i=0}^h \sum_{j=0}^w \omega(i, j) Q_g(i, j) Q_a(i, j)}{\sum_{i=0}^h \sum_{j=0}^w \omega(i, j)} \quad (6)$$

式中, ω 为各像素边界评估的权重图, Q_g 和 Q_a 分别表示参考图像和融合图像边界强度和方向的相似度评估图。该评价指标重点评价与人眼视觉系统相关的图像边界相似性。

其他评价指标还包括视觉信息保真度(VIF)(Han等,2013)、信息传输量(universal quality index, UQI)(Alparone等,2008)和光谱角度相似度(spectral angle mapper, SAM)(Alparone等,2007)等。

2)无参考评价指标。当参考图像不可用时,一些无参考评价准则可对融合图像自身给出不同方面的质量评价。相关的指标包括信息熵(entropy, EN)、梯度域强度(spatial frequency, SF)等。具体公式定义为

$$EN = - \sum_{k=1}^K h_k \log_2 h_k \quad (7)$$

式中, K 代表灰度级个数, h_k 是对应的归一化直方图中的第 k 个分量。该评价指标在一定程度上体现了融合图像所携带的信息量。但是,需要指出的是,

该指标极易受噪声影响。

$$SF = \sqrt{\sum_{i=0}^h \sum_{j=0}^w (\nabla_h I_f(i,j))^2 + \sum_{i=0}^h \sum_{j=0}^w (\nabla_v I_f(i,j))^2} \quad (8)$$

式中, ∇_h 和 ∇_v 分别为水平梯度算子和垂直梯度算子。该评价指标是融合图像边界和纹理信息程度的度量。类似的梯度评价指标还有梯度域平均强度 (average gradient, AG) (Cui 等, 2015)。

最后, 本文认为有必要讨论 1.3 小节损失函数和本小节评价指标的关系。一个合理的损失函数一定能作为评价指标对比不同融合算法的性能; 然而, 一个合理的评价指标却并不一定能胜任损失函数的角色。其根本原因是端对端优化 (或感知优化) 任务的难度比质量评价任务的难度大很多。融合算法在感知优化过程中会生成一系列失真各样的过渡图片, 这便要求损失函数准确捕捉并度量这些失真程度, 从而引导算法最终生成感知质量好的融合图片。评价指标则只需要根据不同融合图像的相对质量对其进行排序, 该过程设计的失真类型、大小相对单一可控。

2 其他图像融合任务及应用

首先引入选择性图像融合问题建模, 并以基于高质量纹理图融合的深度图增强任务为例, 阐述该技术进展。此外, 还阐述了部分基于图像融合的其他视觉任务进展。

2.1 选择性图像融合问题建模

第 1 节考虑了一般性图像融合问题, 即通过融合不同设备、不同设置下获取的多幅图像信息表征优势, 实现高质量的单图像展示。事实上, 图像融合还有一些其他的变体建模, 例如, 选择性图像融合。通常, 当待融合的图像具有不同的物理背景含义或具有语义分歧时, 信息的融合必然具有选择性, 否则融合后的图像将具有明显歧义, 与场景真实情况不符。更进一步, 在多种不同物理背景含义图像之间, 一般图像质量并不均衡, 故一类特殊的选择性图像融合问题还可能具有方向性, 即定向选择性图像融合。一个典型的应用是基于高分辨率纹理图融合的深度图增强。具体地, 彩色纹理图的获取成本较低, 甚至可以使用手机等移动设备拍摄高分辨率的纹理图像。然而, 作为场景几何信息表征的深度图相对

较难获取。一般而言, 虽然同一场景获取的纹理图和深度图在一定层次的语义层面具有相似性, 例如, 物体轮廓, 但由于它们不同的物理背景含义, 导致这两类图像的数据分布流形具有明显区别, 例如, 贴图纹理不应出现在深度图中。

为避免错误的信息融合, 国内外研究者往往设计模型挖掘高质量纹理图的可用信息, 并选择性融入深度图中, 旨在提升深度图质量的同时抑制融合赝像, 包括贴图纹理拷贝和深度图边界模糊等。Hui 等人 (2016) 提出了这类问题的一个基础监督学习框架。框架设计上, 它采用了双向特征提取, 即纹理图通过编码器逐步抽象, 深度图则通过类似解码器过程逐步增强细节。多尺度的深度特征和纹理特征通过直接通道合并后学习特殊卷积核实现选择性特征融合。为进一步提升网络训练鲁棒性, 研究者引入了一些改进, 包括残差学习、批归一化等 (Zuo 等, 2020)。随着注意力机制研究进展, 一些工作通过结合注意力机制, 自适应定向融合上述跨域特征, 这个时期的工作大多基于黑盒端对端模型训练的方式, 故模型解释性方面存在一定挑战, 其限制了性能的进一步提高。

为给模型设计提供明确的方向, 国内外研究者充分回顾了传统模型, 并基于深度神经网络工具重塑经典先验。模型驱动时代, 马尔可夫随机场是该方向的主流选择。所以, 一类网络设计以直接学习马尔可夫随机场数据项和正则项为目标 (Riegler 等, 2016), 该设计结合了传统模型的解释性和神经网络在特征提取方面的优势。其他尝试还包括多尺度卷积核学习 (Wen 等, 2019)。

此外, 站在传统模型视角考虑纹理图和深度图的数据分布流形差异, 一些网络从设计角度深入结合传统模型中抑制不合理纹理拷贝和模糊边界的设计。其主旨是对纹理指导特征进行筛选, 以提升指导特征与相应深度特征的相关性, 例如, 从频率成分分解角度提升高频纹理特征与深度特征的相关性 (He 等, 2021)、从引导性滤波的仿射变换角度强化纹理边界与深度边界的一致性 (Zuo 等, 2021)。

除了采用分量分解方式, 另一条研究分支受传统非线性滤波影响, 例如, 双边滤波器、三边滤波器等, 其致力于从卷积滤波方式上变革, 实现内容敏感的卷积核推断和特征提取。该研究思路的首个工作 (Li 等, 2016) 并没有提供精细化模拟传统三边滤波

器的方法,仍然采用通道合并作为特征融合手段,其仅在网络宏观结构上模拟了三边滤波器。受到该工作的启发,其他代表性工作通过预定义内容感知卷积核生成函数(Su等,2019)、独立推断所有像素滤波器感受野和卷积核(Kim等,2021)等方式模拟传统三边滤波器。

2.2 基于图像融合的其他视觉任务进展

针对各类图像融合任务特点,阐述基于图像融合技术在其他视觉任务的进展。

1) 基于图像融合的低级视觉应用。作为低级视觉任务,图像融合还可促进其他低级视觉任务的发展,例如,图像去噪(Mildenhall等,2018)、图像去模糊(Shang等,2021)、图像超分辨(Bhat等,2021)、图像去雾(Schechner等,2001)以及低光照图像增强(Lu和Zhang,2021)等。这类图像增强问题面临的一个主要挑战是增强后的图像在亮度、色彩表现和对比度等方面与真实图像相比存在一定程度的失真。为解决该缺陷,一些学者显式地将图像增强问题建模成一组图像融合问题。以低光照图像增强为例,相关工作通常生成准多曝光方式采集图像序列,比如使用伽马校正、空间色彩饱和度线性变换等预处理方法。多曝光图像融合可提供高对比度且饱和度均匀的初始增强图像,有利于抑制极端暗部分的信息丢失和边界失真(Zhu等,2020;Lu和Zhang,2021)。

2) 基于图像融合的高级视觉应用。除了上述低级任务中的应用,图像融合对下游计算机视觉任务同样具有指导意义,例如,目标跟踪。传统的目标跟踪方法往往仅针对单模态图像和视频定位感兴趣目标。相应地,基于可见光的目标跟踪技术在夜间和低光照情况下的性能很难得到保障。相反,基于红外成像的目标跟踪技术缺乏充分的纹理信息,因此其性能亦不鲁棒。为解决上述矛盾,研究者提出基于可见光和红外成像融合技术的目标跟踪算法。该研究方向的核心技术之一即为高效的可见光和红外成像融合问题,其显著影响算法的性能(Zhang等,2016;Lan等,2019)。

3 开放挑战与技术展望

深度学习工具的引入对图像融合各类任务起到了明显的推动作用。特别地,基于数据驱动的分层

特征表达和端对端模型训练,相应的深度模型比传统方法表现出显著的性能增益。近年来,一些新的网络结构和训练技巧进一步丰富了深度学习理论的内涵,给图像融合领域持续注入了新的活力。然而,基于深度学习的图像融合算法研究仍面临一些挑战,本文以列举笔者认为重要的挑战结束,期待领域学者们在这些方向上取得重要突破:

1) 从数据集生成角度出发,如何清晰明确地定义不同融合应用的 Ground-truth 融合图像,是大规模公平比较融合算法的前提。在此基础上,如何高效准确生成 Ground-truth 是训练泛化性强的基于深度神经网络的融合算法的保证。进一步地,如何准备源图像序列是进一步提升融合算法性能的关键。

2) 从神经网络构造角度出发,一个有前景的方向是设计轻量的、可解释的和泛化性好的模型组件。如,在高光谱图像融合领域,不同传感器对应不同的观察模型,这便要求融合算法具有跨观察模型泛化性能。如图4所示,现有主流框架分为早期融合和后期融合两种策略。针对不同应用场景的最佳融合策略值得深入研究。最后,在探索单个神经网络实现多个融合任务的过程中,合理地进行参数共享和参数隔离,让不同融合任务相互促进而非相互限制,是追求融合算法统一框架的现实意义所在。

3) 从损失函数设计角度出发,一个亟待验证的问题是现阶段大量使用的对不同损失函数加权求和得到的最终损失函数是否(人类视觉或机器视觉)感知相关。验证方式有很多,如,直接以该损失函数在图像域进行感知优化(Ding等,2021),观察优化中的过渡图像序列,或直接在带有人类视觉(或机器视觉)标签的数据集进行验证。与此同时,设计感知相关的、数学性质良好的以及计算复杂度低的损失函数是图像融合领域的首要(也是终极)任务。这是因为感知最优的融合算法是对该损失函数进行感知优化的直接产物。

4) 从模型训练角度出发,对抗学习现今是图像融合领域的标配。然而,选取哪些类型的图片作为判别器的输入在不同融合应用背景下值得深入思考。再者,在融合算法测试过程中,对其基于测试序列进行微调,有一定可能进一步提升融合性能。该过程在机器学习领域称为测试期间训练(test-time training),已广泛运用于各种图像处理、计算机视觉任务中(Sun等,2020;Mohan等,2021)。

5)从评价指标角度出发,由于现有的评价指标并不能真实有效地反映融合算法的感知表现,如何高效地、去偏地进行主观质量评价(Cao等,2021)需要重新得到重视。

参考文献 (References)

- Alparone L, Aiazzi B, Baronti S, Garzelli A, Nencini F and Selva M. 2008. Multispectral and panchromatic data fusion assessment without reference. *Photogrammetric Engineering and Remote Sensing*, 74(2): 193-200 [DOI: 10.14358/PERS.74.2.193]
- Alparone L, Wald L, Chanussot J, Thomas C, Gamba P and Bruce L M. 2007. Comparison of pansharpening algorithms: outcome of the 2006 GRS-S data-fusion contest. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10): 3012-3021 [DOI: 10.1109/TGRS.2007.904923]
- Ballester C, Caselles V, Igual L, Verdera J and Rougé 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]
- Barron J T, Adams A, Shih Y and Hernández C. 2015. Fast bilateral-space stereo for synthetic defocus//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA; IEEE: 4466-4474 [DOI: 10.1109/CVPR.2015.7299076]
- Bhat G, Danelljan M, Van Gool L and Timofte R. 2021. Deep burst super-resolution//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, USA; IEEE: 9205-9214 [DOI: 10.1109/CVPR46437.2021.00909]
- Bhatnagar G, Wu Q M J and Liu Z. 2013. Directive contrast based multimodal medical image fusion in NSCT domain. *IEEE Transactions on Multimedia*, 15(5): 1014-1024 [DOI: 10.1109/TMM.2013.2244870]
- Cai J R, Gu S H and Zhang L. 2018. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4): 2049-2062 [DOI: 10.1109/TIP.2018.2794218]
- Cao P B, Wang Z Y and Ma K D. 2021. Debaised subjective assessment of real-world image enhancement//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, USA; IEEE: 711-721 [DOI: 10.1109/CVPR46437.2021.00077]
- Cui G M, Feng H J, Xu Z H, Li Q and Chen Y T. 2015. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341: 199-209 [DOI: 10.1016/j.optcom.2014.12.032]
- Deng X and Dragotti P L. 2021. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10): 3333-3348 [DOI: 10.1109/TPAMI.2020.2984244]
- Ding K Y, Ma K D, Wang S Q and Simoncelli E P. 2021. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129(4): 1258-1281 [DOI: 10.1007/s11263-020-01419-7]
- Ding X H, Zhang X Y, Zhou Y Z, Han J G, Ding G G and Sun J. 2022. Scaling up your kernels to 31×31 : revisiting large kernel design in CNNs[EB/OL]. [2022-06-28]. <http://arxiv.org/pdf/2203.06717.pdf>
- Di Zeno S. 1986. A note on the gradient of a multi-image. *Computer Vision, Graphics, and Image Processing*, 33(1): 116-125 [DOI: 10.1016/0734-189X(86)90223-9]
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Housby N. 2021. An image is worth 16×16 words: transformers for image recognition at scale [EB/OL]. [2022-06-28]. <http://arxiv.org/pdf/2010.11929.pdf>
- Dudhane A, Zamir S W, Khan S, Khan F S and Yang M H. 2022. Burst image restoration and enhancement [EB/OL]. [2022-04-14] <https://arxiv.org/pdf/2110.03680.pdf>
- Fu Y, Zhang T, Zheng Y Q, Zhang D B and Huang H. 2019. Hyper-spectral image super-resolution with optimized RGB guidance//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, USA; IEEE: 11653-11662 [DOI: 10.1109/CVPR.2019.01193]
- Ghassemian H. 2016. A review of remote sensing image fusion methods. *Information Fusion*, 32: 75-89 [DOI: 10.1016/j.inffus.2016.03.003]
- Guo X P, Nie R C, Cao J D, Zhou D M, Mei L Y and He K J. 2019. FuseGAN: learning to fuse multi-focus image via conditional generative adversarial network. *IEEE Transactions on Multimedia*, 21(8): 1982-1996 [DOI: 10.1109/TMM.2019.2895292]
- Gupta M, Iso D and Nayar S K. 2013. Fibonacci exposure bracketing for high dynamic range imaging//*Proceedings of 2013 IEEE International Conference on Computer Vision*. Sydney, Australia; IEEE: 1473-1480 [DOI: 10.1109/ICCV.2013.186]
- Han Y, Cai Y Z, Cao Y and Xu X M. 2013. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2): 127-135 [DOI: 10.1016/j.inffus.2011.08.002]
- He L Z, Zhu H G, Li F, Bai H H, Cong R M, Zhang C J, Lin C Y, Liu M Q and Zhao Y. 2021. Towards fast and accurate real-world depth super-resolution: benchmark dataset and baseline//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, USA; IEEE: 9225-9234 [DOI: 10.1109/CVPR46437.2021.00911]
- Hermessi H, Mourali O and Zagrouba E. 2021. Multimodal medical image fusion review: theoretical background and recent advances. *Signal Processing*, 183: #108036 [DOI: 10.1016/j.sigpro.2021.108036]
- Hu J W, Hu P, Kang X D, Zhang H and Fan S S. 2021. Pan-sharpening via multiscale dynamic convolutional neural network. *IEEE*

- Transactions on Geoscience and Remote Sensing, 59(3): 2231-2244 [DOI: 10.1109/TGRS.2020.3007884]
- Hui T W, Loy C C and Tang X O. 2016. Depth map super-resolution by deep multi-scale guidance//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 353-369 [DOI: 10.1007/978-3-319-46487-9_22]
- Jin S P, Yu B B, Jing M H, Zhou Y, Liang J J and Ji R H. 2022. DarkVisionNet: low-light imaging via RGB-NIR fusion with deep inconsistency prior. Proceedings of the AAAI Conference on Artificial Intelligence, 36(1): 1104-1112 [DOI: 10.1609/aaai.v36i1.19995]
- Jin X, Wang P and Dai Q H. 2020. Parallax tolerant light field stitching for hand-held plenoptic cameras. IEEE Transactions on Image Processing, 29: 1929-1943 [DOI: 10.1109/TIP.2019.2945687]
- Jung H, Kim Y, Jang H, Ha N and Sohn K. 2020. Unsupervised deep image fusion with structure tensor representations. IEEE Transactions on Image Processing, 29: 3845-3858 [DOI: 10.1109/TIP.2020.2966075]
- Kim B, Ponce J and Ham B. 2021. Deformable kernel networks for joint image filtering. International Journal of Computer Vision, 129(2): 579-600 [DOI: 10.1007/s11263-020-01386-z]
- Lan X Y, Ye M, Shao R, Zhong B N, Yuen P C and Zhou H Y. 2019. Learning modality-consistency feature templates: a robust RGB-infrared tracking system. IEEE Transactions on Industrial Electronics, 66(12): 9887-9897 [DOI: 10.1109/TIE.2019.2898618]
- Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z and Shi W. 2017. Photo-realistic single image super-resolution using a generative adversarial network//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 105-114 [DOI:10.1109/CVPR.2017.19]
- Li H and Wu X J. 2019. DenseFuse: a fusion approach to infrared and visible images. IEEE Transactions on Image Processing, 28(5): 2614-2623 [DOI: 10.1109/TIP.2018.2887342]
- Li H, Wu X J and Kittler J. 2020a. MDLatLRR: a novel decomposition method for infrared and visible image fusion. IEEE Transactions on Image Processing, 29: 4733-4746 [DOI: 10.1109/TIP.2020.2975984]
- Li H F, Cen Y L, Liu Y, Chen X and Yu Z T. 2021. Different input resolutions and arbitrary output resolution: a meta learning-based deep framework for infrared and visible image fusion. IEEE Transactions on Image Processing, 30: 4070-4083 [DOI: 10.1109/TIP.2021.3069339]
- Li J X, Guo X B, Lu G M, Zhang B, Xu Y, Wu F and Zhang D. 2020b. DRPL: deep regression pair learning for multi-focus image fusion. IEEE Transactions on Image Processing, 29: 4816-4831 [DOI: 10.1109/TIP.2020.2976190]
- Li S T, Kang X D and Hu J W. 2013. Image fusion with guided filtering. IEEE Transactions on Image Processing, 22(7): 2864-2875 [DOI: 10.1109/TIP.2013.2244222]
- Li Y J, Huang J B, Ahuja N and Yang M H. 2016. Deep joint image filtering//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 154-169 [DOI: 10.1007/978-3-319-46493-0_10]
- Liu J Y, Fan X, Huang Z B, Wu G Y, Liu R S, Zhong W and Luo Z X. 2022. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 5792-5801 [DOI: 10.1109/CVPR52688.2022.00571]
- Liu M W, Wang R H, Li J and Jiao Y Z. 2021. Infrared and visible image fusion with multi-scale anisotropic guided filtering. Journal of Image and Graphics, 26(10): 2421-2432 (刘明葳, 王任华, 李静, 焦映臻. 2021. 各向异性导向滤波的红外与可见光图像融合. 中国图象图形学报, 26(10): 2421-2432) [DOI: 10.11834/jig.200339]
- Liu Q J, Zhou H Y, Xu Q Z, Liu X Y and Wang Y H. 2021a. PSGAN: a generative adversarial network for remote sensing image pan-sharpening. IEEE Transactions on Geoscience and Remote Sensing, 59(12): 10227-10242 [DOI: 10.1109/TGRS.2020.3042974]
- Liu Y, Chen X, Peng H and Wang Z F. 2017. Multi-focus image fusion with a deep convolutional neural network. Information Fusion, 36: 191-207 [DOI: 10.1016/j.inffus.2016.12.001]
- Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021b. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]
- Loshchilov I and Hutter F. 2017. SGDR: stochastic gradient descent with warm restarts [EB/OL]. [2022-05-20]. <http://arxiv.org/pdf/1608.03983.pdf>
- Lu K and Zhang L H. 2021. TBEFN: a two-branch exposure-fusion network for low-light image enhancement. IEEE Transactions on Multimedia, 23: 4093-4105 [DOI: 10.1109/TMM.2020.3037526]
- Ma J Y, Tang L F, Fan F, Huang J, Mei X G and Ma Y. 2022. Swin-Fusion: cross-domain long-range learning for general image fusion via Swin transformer. IEEE/CAA Journal of Automatica Sinica, 9(7): 1200-1217 [DOI: 10.1109/JAS.2022.105686]
- Ma J Y, Xu H, Jiang J J, Mei X G and Zhang X P. 2020a. DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Transactions on Image Processing, 29: 4980-4995 [DOI: 10.1109/TIP.2020.2977573]
- Ma J Y, Yu W, Chen C, Liang P W, Guo X J and Jiang J J. 2020b. 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 K D, Duanmu Z F, Yeganeh H and Wang Z. 2018. Multi-exposure image fusion by optimizing a structural similarity index. *IEEE Transactions on Computational Imaging*, 4(1): 60-72 [DOI: 10.1109/TCI.2017.2786138]
- Ma K D, Duanmu Z F, Zhu H W, Fang Y M and Wang Z. 2020c. Deep guided learning for fast multi-exposure image fusion. *IEEE Transactions on Image Processing*, 29: 2808-2819 [DOI: 10.1109/TIP.2019.2952716]
- Meng X C, Shen H F, Yuan Q Q, Li H F, Zhang L P and Sun W W. 2019. Pansharpening for cloud-contaminated very high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(5): 2840-2854 [DOI: 10.1109/TGRS.2018.2878007]
- Mildenhall B, Barron J T, Chen J W, Sharlet D, Ng R and Carroll R. 2018. Burst denoising with kernel prediction networks//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA; IEEE: 2502-2510 [DOI: 10.1109/CVPR.2018.00265]
- Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N and Terzopoulos D. 2022. Image segmentation using deep learning: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3523-3542 [DOI: 10.1109/TPAMI.2021.3059968]
- Mohan S, Vincent J L, Manzorro R, Crozier P, Fernandez-Granda C and Simoncelli E. 2021. Adaptive denoising via gaintuning//*Proceedings of the 34th International Conference on Neural Information Processing Systems*. [s.l.]: Curran Associates, Inc.: 23727-23740
- Park N and Kim S. 2022. How do vision transformers work? [EB/OL]. [2022-06-28]. <http://arxiv.org/pdf/2202.06709.pdf>
- Petschnigg G, Szeliski R, Agrawala M, Cohen M, Hoppe H and Toyama K. 2004. Digital photography with flash and no-flash image pairs. *ACM Transactions on Graphics*, 23(3): 664-672 [DOI: 10.1145/1015706.1015777]
- Prabhakar K R, Srikanth V S and Babu R V. 2017. DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs//*Proceedings of the 2017 IEEE International Conference on Computer Vision*. Venice, Italy; IEEE: 4724-4732 [DOI: 10.1109/ICCV.2017.505]
- Qu L H, Liu S L, Wang M N and Song Z J. 2022. TransMEF: a transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2): 2126-2134 [DOI: 10.1609/aaai.v36i2.20109]
- Qu Y, Qi H R and Kwan C. 2018. Unsupervised sparse dirichlet-net for hyperspectral image super-resolution//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA; IEEE: 2511-2520 [DOI: 10.1109/CVPR.2018.00266]
- Rahman H, Soundararajan R and Babu R V. 2017. Evaluating multiexposure fusion using image information. *IEEE Signal Processing Letters*, 24(11): 1671-1675 [DOI: 10.1109/LSP.2017.2752233]
- Riegler G, Ferstl D, R  ther M and Bischof H. 2016. A deep primal-dual network for guided depth super-resolution [EB/OL]. [2022-05-26]. <http://arxiv.org/pdf/1607.08569.pdf>
- Schechner Y Y, Narasimhan S G and Nayar S K. 2001. Instant dehazing of images using polarization//*Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Kauai, USA; IEEE: 325-332 [DOI: 10.1109/CVPR.2001.990493]
- Shang W, Ren D W, Zou D Q, Ren J S, Luo P and Zuo W M. 2021. Bringing events into video deblurring with non-consecutively blurry frames//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*. Montreal, Canada; IEEE: 4511-4520 [DOI: 10.1109/ICCV48922.2021.00449]
- Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2022-09-21]. <http://arxiv.org/pdf/1409.1556.pdf>
- Son C H and Zhang X P. 2016. Layer-based approach for image pair fusion. *IEEE Transactions on Image Processing*, 25(6): 2866-2881 [DOI: 10.1109/TIP.2016.2556618]
- Su H, Jampani V, Sun D Q, Gallo O, Learned-Miller E and Kautz J. 2019. Pixel-adaptive convolutional neural networks//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, USA; IEEE: 11158-11167 [DOI: 10.1109/CVPR.2019.01142]
- Su X Y, Li J J and Hua Z. 2022. Transformer-based regression network for pansharpening remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: #5407423 [DOI: 10.1109/TGRS.2022.3152425]
- Sun Y, Wang X L, Liu Z, Miller J, Efros A A and Hardt M. 2020. Test-time training with self-supervision for generalization under distribution shifts//*Proceedings of the 37th International Conference on Machine Learning*. [s.l.]: JMLR.org: 9229-9248
- Tang L F, Yuan J T and Ma J Y. 2022a. Image fusion in the loop of high-level vision tasks: a semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28-42 [DOI: 10.1016/j.inffus.2021.12.004]
- Tang L F, Yuan J T, Zhang H, Jiang X Y and Ma J Y. 2022b. PIAFusion: a progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84: 79-92 [DOI: 10.1016/j.inffus.2022.03.007]
- Vivone G. 2019. Robust band-dependent spatial-detail approaches for panchromatic sharpening. *IEEE transactions on Geoscience and Remote Sensing*, 57(9): 6421-6433 [DOI: 10.1109/TGRS.2019.2906073]
- Wang B, Niu H F, Zeng J C, Bai G F, Lin S Z and Wang Y B. 2021. Latent representation learning model for multi-band images fusion via

- low-rank and sparse embedding. *IEEE Transactions on Multimedia*, 23: 3137-3152 [DOI: 10.1109/TMM.2020.3020695]
- Wang L, Li B and Tian L F. 2014. EGGDD: an explicit dependency model for multi-modal medical image fusion in shift-invariant shearlet transform domain. *Information fusion*, 19: 29-37 [DOI: 10.1016/j.inffus.2013.04.005]
- Wang L F, Dou J L, Qin P L, Lin S Z, Gao Y and Zhang C C. 2019. Medical image fusion using double dictionary learning and adaptive PCNN. *Journal of Image and Graphics*, 24(9): 1588-1603 (王丽芳, 窦杰亮, 秦品乐, 蔺素珍, 高媛, 张程程. 2019. 双重字典学习与自适应 PCNN 相结合的医学图像融合. *中国图象图形学报*, 24(9): 1588-1603) [DOI: 10.11834/jig.180667]
- Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600-612 [DOI: 10.1109/TIP.2003.819861]
- Wang Z X, Zhang J W, Lin M D, Wang J, Luo P and Ren J. 2020. Learning a reinforced agent for flexible exposure bracketing selection//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA: IEEE: 1817-1825 [DOI: 10.1109/CVPR42600.2020.00189]
- Wei Q, Bioucas-Dias J, Dobigeon N and Tourneret J Y. 2015. Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7): 3658-3668 [DOI: 10.1109/TGRS.2014.2381272]
- Wen Y, Sheng B, Li P, Lin W Y and Feng D D. 2019. Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution. *IEEE Transactions on Image Processing*, 28(2): 994-1006 [DOI: 10.1109/TIP.2018.2874285]
- Xiao Y, Yang M M, Li C L, Liu L and Tang J. 2022. Attribute-based progressive fusion network for RGBT tracking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3): 2831-2838 [DOI: 10.1609/aaai.v36i3.20187]
- Xie Q, Zhou M H, Zhao Q, Xu Z B and Meng D Y. 2022. MHF-net: an interpretable deep network for multispectral and hyperspectral image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3): 1457-1473 [DOI: 10.1109/TPAMI.2020.3015691]
- Xie W Y, Cui Y H, Li Y S, Lei J, Du Q and Li J J. 2021. HPGAN: hyperspectral pansharpening using 3-D generative adversarial networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1): 463-477 [DOI: 10.1109/TGRS.2020.2994238]
- Xu H and Ma J Y. 2021. EMFusion: an unsupervised enhanced medical image fusion network. *Information Fusion*, 76: 177-186 [DOI: 10.1016/j.inffus.2021.06.001]
- Xu H, Ma J Y, Jiang J J, Guo X J and Ling H B. 2022a. U2Fusion: a unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502-518 [DOI: 10.1109/TPAMI.2020.3012548]
- Xu H, Ma J Y, Shao Z F, Zhang H, Jiang J J and Guo X J. 2021. SDP-Net: 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]
- Xu H, Ma J Y, Yuan J T, Le Z L and Liu W. 2022b. RFNet: unsupervised network for mutually reinforcing multi-modal image registration and fusion//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, USA: IEEE: 19647-19656 [DOI: 10.1109/CVPR52688.2022.01906]
- Xu S, Amira O, Liu J M, Zhang C X, Zhang J S and Li G H. 2020. HAM-MFN: hyperspectral and multispectral image multiscale fusion network with RAP loss. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7): 4618-4628 [DOI: 10.1109/TGRS.2020.2964777]
- Yang J F, Fu X Y, Hu Y W, Huang Y, Ding X H and Paisley J. 2017. PanNet: a deep network architecture for pan-sharpening//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Venice, Italy: IEEE: 1753-1761 [DOI: 10.1109/ICCV.2017.193]
- Yang S Y, Wang M, Jiao L C, Wu R X and Wang Z X. 2010. Image fusion based on a new contourlet packet. *Information Fusion*, 11(2): 78-84 [DOI: 10.1016/j.inffus.2009.05.001]
- Yao J, Hong D F, Chanussot J, Meng D Y, Zhu X X and Xu Z B. 2020. Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution//*Proceedings of the 16th European Conference on Computer Vision*. Glasgow, UK: Springer: 208-224 [DOI: 10.1007/978-3-030-58526-6_13]
- Ye X C, Sun B L, Wang Z H, Yang J Y, Xu R, Li H J and Li B P. 2020. PMBANet: progressive multi-branch aggregation network for scene depth super-resolution. *IEEE Transactions on Image Processing*, 29: 7427-7442 [DOI: 10.1109/TIP.2020.3002664]
- Yokoya N, Yairi T and Iwasaki A. 2012. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2): 528-537 [DOI: 10.1109/TGRS.2011.2161320]
- Zhang H and Ma J Y. 2021a. SDNet: a versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10): 2761-2785 [DOI: 10.1007/s11263-021-01501-8]
- Zhang H, Xu H, Xiao Y, Guo X J and Ma J Y. 2020a. Rethinking the image fusion: a fast unified image fusion network based on proportional maintenance of gradient and intensity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7): 12797-12804 [DOI: 10.1609/aaai.v34i07.6975]
- Zhang K H, Liu Q S, Wu Y and Yang M H. 2016. Robust visual tracking via convolutional networks without training. *IEEE Transactions on Image Processing*, 25(4): 1779-1792 [DOI: 10.1109/TIP.2016.2531283]
- Zhang X T, Huang W, Wang Q and Li X L. 2021b. SSR-NET: spatial

- spectral reconstruction network for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7): 5953-5965 [DOI: 10.1109/TGRS.2020.3018732]
- Zhang Y, Liu Y, Sun P, Yan H, Zhao X L and Zhang L. 2020b. IFC-NN: a general image fusion framework based on convolutional neural network. *Information Fusion*, 54: 99-118 [DOI: 10.1016/j.inffus.2019.07.011]
- Zhao W D, Wang D and Lu H C. 2019. Multi-focus image fusion with a natural enhancement via a joint multi-level deeply supervised convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(4): 1102-1115 [DOI: 10.1109/TCSVT.2018.2821177]
- Zhao Z X, Xu S, Zhang C X, Liu J M, Zhang J S and Li P F. 2021. DIDFuse: deep image decomposition for infrared and visible image fusion//*Proceedings of the 29th International Joint Conference on Artificial Intelligence*. Yokohama, Japan; [s. n.]: 970-976 [DOI: 10.24963/ijcai.2020/135]
- Zheng K, Gao L R, Liao W Z, Hong D F, Zhang B, Cui X M and Channusot J. 2021. Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3): 2487-2502 [DOI: 10.1109/TGRS.2020.3006534]
- Zheng M Y, Qi G Q, Zhu Z Q, Li Y Y, Wei H Y and Liu Y. 2020. Image dehazing by an artificial image fusion method based on adaptive structure decomposition. *IEEE Sensors Journal*, 20(14): 8062-8072 [DOI: 10.1109/JSEN.2020.2981719]
- Zhu M F, Pan P B, Chen W and Yang Y. 2020. EEMEFN: low-light image enhancement via edge-enhanced multi-exposure fusion network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7): 13106-13113 [DOI: 10.1609/aaai.v34i07.7013]
- Zhu X X and Bamler R. 2013. A sparse image fusion algorithm with application to pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 51(5): 2827-2836 [DOI: 10.1109/TGRS.2012.2213604]
- Zhu Y B, Li C L, Luo B, Tang J and Wang X. 2019. Dense feature aggregation and pruning for RGBT tracking//*Proceedings of the 27th ACM International Conference on Multimedia*. Nice, France; ACM: 465-472 [DOI: 10.1145/3343031.3350928]
- Zuo Y F, Fang Y M, An P, Shang X W and Yang J N. 2021. Frequency-dependent depth map enhancement via iterative depth-guided affine transformation and intensity-guided refinement. *IEEE Transactions on Multimedia*, 23: 772-783 [DOI: 10.1109/TMM.2020.2987706]
- Zuo Y F, Wu Q, Fang Y M, An P, Huang L Q and Chen Z F. 2020. Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2): 297-306 [DOI: 10.1109/TCSVT.2018.2890271]

作者简介

左一帆,男,副教授,主要研究方向为计算机视觉、机器学习。

E-mail:kenny0410@126.com

方玉明,通信作者,男,教授,主要研究方向为人工智能、多媒体大数据处理。E-mail:leo.fangyuming@foxmail.com

马柯德,男,助理教授,主要研究方向为多媒体内容质量评与感知优化。E-mail:kede.ma@cityu.edu.hk