**引用**:朱金鹏,姜文博,杜涛. 基于 Mamba 模型的可见光-红外-偏振图像融合方法[J]. 聊城大学学报(自然科学版),2026,39(6):823-832.

**Citation**: Zhu Jinpeng, Jiang Wenbo, Du Tao. Visible-infrared-polarization image fusion based on Mamba model[J]. Journal of Liaocheng University (Natural Science Edition), 2026, 39(6): 823-832.

**文章编号** 1672-6634(2026)06-0823-10 **DOI** 10.19728/j.issn1672-6634.2026040009

# 基于 Mamba 模型的可见光-红外-偏振图像融合方法

朱金鹏,姜文博,杜涛

(北方工业大学 人工智能与计算机学院,北京 100144)

**摘要** 针对低照度和弱纹理等复杂退化环境导致单模态视觉传感器感知受限等问题,提出了一种面向低光场景的可见光-红外-偏振三模态协同感知与融合网络。首先,构建多分支特征提取架构,引入状态空间模型(Mamba)实现异构特征的全局长程依赖建模;其次,采用以可见光纹理为锚点的双路交互融合策略,实现多模态信息的深度互补;最后,设计自适应亮度保真与局部梯度感知的复合损失函数。实验表明,该方法在信息熵等指标上取得最优表现,在有效保留偏振高频特征与可见光纹理的同时,显著提升了红外热辐射的全局对比度,增强了低光环境下的多传感器透彻感知。

**关键词** 图像融合;偏振成像;多传感器感知;Mamba 模型

**中图分类号** TP391.4 **文献标志码** A

# Visible-infrared-polarization image fusion based on Mamba model

ZHU Jinpeng, JIANG Wenbo, DU Tao

(School of Artificial Intelligence and Computer Science, North China University of Technology, Beijing 100144, China)

**Abstract** To address limitations in single-modal visual sensor perception under low-light and weak-texture conditions, this paper proposes a visible-light-infrared-polarization tri-modal collaborative perception and fusion network tailored for low-light scenarios. First, a multi-branch feature extraction architecture is constructed, incorporating the State Space Model (Mamba) to enable full-length long-range dependency modeling of heterogeneous features. Second, a dual-path interactive fusion strategy using visible light textures as anchors achieves deep multimodal complementarity. Finally, an adaptive composite loss function combining brightness fidelity preservation and local gradient awareness is designed. Experimental results demonstrate that this approach achieves optimal performance metrics such as information entropy, effectively retaining polarized high-frequency features and visible light textures while enhancing global contrast of infrared thermal radiation, thereby enabling comprehensive multi-sensor perception in low-light environments.

**Keywords** image fusion; polarization imaging; multi-sensor perception; Mamba model

**收稿日期**:2026-04-20

**基金项目**:国家重点研发计划项目(2023YFB4704404);国家自然科学基金项目(62388101)资助

**通信作者**:杜涛,男,汉族,博士,副教授,博士生导师,研究方向:智能导航与信息融合,E-mail:tdu@ncut.edu.cn。

# 0 引言

在低照度、强噪声和杂光干扰等复杂光照环境中,单模态成像面临信噪比下降、特征维数受限等瓶颈,难以满足高精度目标检测、识别与导航等任务的鲁棒性需求。为突破传统单模态成像在信息获取维度上的极限,多模态图像融合已成为图像融合领域的前沿热点[1-2]。在现有的视觉感知方式中,可见光成像可以高保真地刻画场景的精细纹理与结构轮廓,但在低光照或非均匀光照条件下,其细节表征能力往往大幅退化。红外成像则基于目标与背景的热辐射差异,具备优异的穿透性和目标突显能力,在弱光环境中优势显著。然而,红外图像普遍存在空间分辨率受限、纹理细节缺失等问题。仅仅依赖可见光-红外的融合虽能实现一定程度的优势互补,但本质上可见光与红外融合仍局限于光强标量信息的叠加,难以从物理层面上滤除强杂光干扰或有效辨识材质属性[3]。

与可见光及红外成像均不同,偏振成像利用光波的偏振特性获取信息,通过解算光的偏振状态,可以表征物体表面的材质特性、粗糙度以及几何结构信息,在抑制表面高光反射(耀光干扰)、增强低对比度目标方面展现出独特优势[4]。偏振信息能够发现传统可见光图像中难以获取的隐匿细节,例如被强反射掩盖的目标轮廓。因此,偏振成像在目标检测[5-6]、语义分割[7]以及医疗成像[8]等领域得到了广泛研究。

综上所述,可见光图像、红外图像与偏振图像在信息表征上具备兼容互补特性。可见光图像提供空间纹理,红外图像突显高对比度热辐射特征,而偏振图像则揭示物理属性并抑制杂光。将可见光图像、红外图像和偏振图像深度融合,不仅可以保留场景结构细节,还可以在复杂干扰下增强目标区域的显著性,实现多维光场的多要素联合表征,从而为后续视觉任务提供稳定、可靠的数据底座[9]。

在技术实现层面,现有多模态图像融合方法主要分为传统方法与基于深度学习的方法两类[10]。传统方法依赖数学模型与人工先验规则,如基于多尺度变换[11-12]、稀疏表示方法[13]以及伪彩色编码[14]等。这类方法在静态或受限场景下具备一定有效性,但在面对复杂场景中非独立、异构分布的噪声干扰时,固定的融合规则往往难以自适应新的数据特征,泛化性能显著受限[15]。

随着深度学习技术的发展,数据驱动的融合网络逐渐成为主流研究方向[16]。代表性架构包括基于自动编码器(Autoencoder,AE)[17-18]、卷积神经网络(Convolutional Neural Network, CNN)[19-20]、生成对抗网络(Generative Adversarial Network, GAN)[21-22]以及 Transformer 类模型[23-25]。这些方法在特征提取与融合性能上取得了长足进步,但仍存在固有局限:CNN 受限于局部感受野,难以捕捉全局上下文;GAN 在对抗训练中易陷入模式崩溃与不稳定;Transformer 虽具备卓越的全局长距离建模能力,但其自注意力机制的二次计算复杂度,极大地限制了其在处理高分辨率多模态数据及资源受限设备上的应用。

为了实现全局信息建模能力的同时兼顾计算效率,近年来状态空间模型(State Space Model,SSM)逐渐受到关注[26-28]。特别是 Mamba 架构,通过引入选择性扫描机制,在保持长距离依赖建模能力的同时,实现了随序列长度线性增长的计算复杂度,在性能与效率之间取得了极佳的平衡[29]。这一特性使其在处理高分辨率、多模态异构感知数据时展现出良好的应用潜力。鉴于此,本文以 Mamba 模型为核心架构,研究可见光-红外-偏振三模态图像的高效融合机制,旨在克服低光照条件约束,提升视觉融合感知质量。

# 1 Mamba 模型

近年来,深度学习模型在视觉感知任务中取得了突破性进展。基于 Transformer 架构的模型凭借其强大的全局上下文建模能力,已成为处理复杂视觉场景的主流方案。然而,Transformer 核心的自注意力机制在处理长序列数据或高分辨率图像时,其计算复杂度与内存消耗随输入规模呈平方级增长。这一瓶颈限制了其在资源受限设备或高实时性感知任务中的应用。

针对上述局限性,研究者开始关注 SSM 在长序列建模中的潜力。SSM 能够以严格的线性计算复杂度实现对全局信息的有效捕捉。在此基础上,通过引入选择性扫描机制的 Mamba 架构,不仅完美继承了循环神经网络的高效递推推理特性,更展现出与 Transformer 相当的全局建模表征能力,为高效的视觉感知任务提供了一种新的解决思路。

状态空间模型是一类用于描述动态系统内部状态演化规律及其输入-输出映射关系的数学模型。在连续时间域下，对于一维输入信号 $\boldsymbol{x}(t)\in\mathbf{R}^1$，系统通过隐状态向量 $\boldsymbol{h}(t)\in\mathbf{R}^n$ 进行演化并生成输出信号 $\boldsymbol{y}(t)\in\mathbf{R}^1$。该过程由如下线性常微分方程组描述[28]

$$\begin{cases}\boldsymbol{h}'(t)=\boldsymbol{A}\boldsymbol{h}(t)+\boldsymbol{B}\boldsymbol{x}(t),\\ \boldsymbol{y}(t)=\boldsymbol{C}\boldsymbol{h}(t)+\boldsymbol{D}\boldsymbol{x}(t),\end{cases}\tag{1}$$

式中 $\boldsymbol{h}(t)$ 是状态向量，$\boldsymbol{h}'(t)$ 为状态向量 $\boldsymbol{h}(t)$ 对时间的一阶导数，$\boldsymbol{x}(t)$ 是输入向量，$\boldsymbol{y}(t)$ 是输出向量，$\boldsymbol{A}\in\mathbf{R}^{n\times n}$ 表示状态转移矩阵，$\boldsymbol{B}\in\mathbf{R}^{n\times 1}$ 为输入映射矩阵，$\boldsymbol{C}\in\mathbf{R}^{1\times n}$ 为状态到输出的投影矩阵，$\boldsymbol{D}\in\mathbf{R}^{1\times 1}$ 表示输入到输出的直接连接项。

在深度学习的实际应用场景中，输入信号通常是离散序列。为了将连续时间模型应用于离散数据处理，必须对系统(1)进行离散化。设定采样时间步长参数 $\Delta t$，采用零阶保持(Zero-Order Hold，ZOH)或双线性变换等离散化方法，将连续状态方程(1)离散化为

$$\begin{cases}\boldsymbol{h}_k=\bar{\boldsymbol{A}}\boldsymbol{h}_{k-1}+\bar{\boldsymbol{B}}\boldsymbol{x}_k,\\ \boldsymbol{y}_k=\boldsymbol{C}\boldsymbol{h}_k+\boldsymbol{D}\boldsymbol{x}_k,\end{cases}\tag{2}$$

式中 $k$ 表示离散时间步，$\bar{\boldsymbol{A}}$ 与 $\bar{\boldsymbol{B}}$ 为离散化后的系统参数矩阵。该递推结构表明，SSM 能够通过对序列进行逐步线性扫描来实现信息传递。其核心优势在于，在推理阶段模型展现出循环神经网络的高效递推特性，计算复杂度随序列长度呈线性增长，这在理论上极大地缓解了长序列建模时的计算压力。

尽管传统 SSM 在序列处理效率上表现优异，但其本质上属于线性时不变(Linear Time-Invariant，LTI)系统。线性假设导致模型缺乏对输入数据的自适应调节能力，难以捕捉复杂的非线性模式与动态演化的语义特征。为弥补这一缺陷，Mamba 架构引入了选择性扫描机制。该机制通过将参数矩阵 $\bar{\boldsymbol{B}}$、$\boldsymbol{C}$ 以及步长 $\Delta t$ 设置为输入 $\boldsymbol{x}_k$ 的函数，使得模型能够根据当前上下文自适应地筛选、压缩或遗忘信息。这种数据驱动的动态参数化策略彻底打破了 LTI 系统的约束，赋予了模型类 Transformer 的内容感知能力，从而构建出更具表达深度的选择性状态空间模型(Selective SSM)。

从宏观拓扑结构分析，Mamba 架构遵循高度模块化的设计理念，其核心计算单元为 Mamba Block。如图 1 所示，该模块深度整合了归一化层、双分支映射结构、状态空间建模算子、门控调制机制及残差融合路径。各子模块在功能上相互独立而在逻辑上协同，旨在理论层面突破 CNN 局部感受野的限制以及 Transformer 架构在计算资源上的瓶颈。这种动态建模机制为处理具有复杂空间依赖性的红外-偏振视觉融合任务提供了稳健的底层特征支撑。

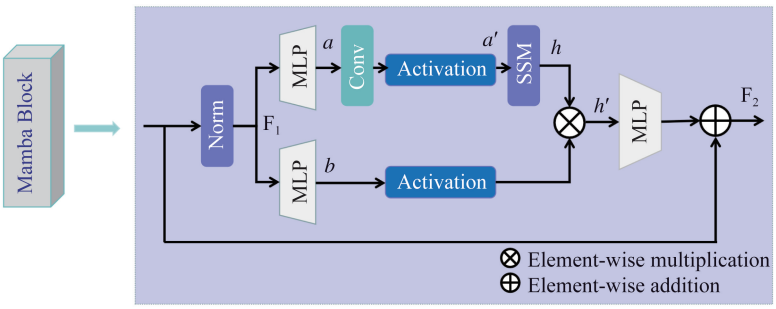



**图 1　Mamba 块**

**Fig. 1　Mamba Block**

## 2　数据集及多源图像预处理

可见光成像虽具备较高的空间分辨率与纹理刻画能力，但其对外部光场依赖程度极高，在弱光环境下不仅信噪比骤降，且极易受阴影与耀光干扰；偏振成像作为光波矢量的表征，能够敏锐捕捉目标材质属性与三维形貌，但在极端低光照条件下，偏振信号的信噪比急剧恶化，导致特征提取困难；红外成像凭借捕获目标与背景的热辐射差异，虽然不受光照影响，但其成像机制决定了其缺乏几何轮廓与纹理特征。

因此,将可见光、红外与偏振三者进行融合,是突破低照度场景感知性能瓶颈的必然路径。然而,与常规的双模态融合不同,三模态联合建模在实际工程中面临两大突出问题:一是不同模态在成像机理和信息表达上存在显著差异,直接拼接或简单叠加容易引入冗余甚至冲突信息。二是目前红外、偏振与可见光三模态严格对齐的数据资源极为有限,公开可用数据规模较小,这对模型结构设计与训练策略提出了更高要求。

本文选取公开数据集中的500个独立场景(每场景包含可见光、红外、偏振三模态图像,共计1 500幅)作为研究基础[16]。按70%、10%、20%随机划分为训练、验证与测试集。考虑到Mamba模型对长程依赖的建模能力依赖于足够的数据多样性,当前数据量仍可能存在过拟合风险,后续将通过数据增强与扩大采集规模进一步验证模型泛化性。鉴于严格空间对齐的三模态开源数据稀缺,未来将搭建专用三模态采集设备,开展严格的几何与辐射标定,构建更大规模、场景更丰富的三模态数据集,以进一步验证和提升方法的泛化能力。经数据分析发现,受限于多传感器采集系统的硬件视场角差异,原始红外图像边缘存在明显的无效盲区(视场边缘伪影及条带状像素缺失)。为避免空间位置失调对后续网络特征提取造成干扰,确保多源异构特征在空间坐标系下的严格像素级对齐,本文在预处理阶段对三类模态图像执行了高精度的中心裁剪与配准。处理后,三模态图像在物理尺寸与感受野上实现了统一标准化,典型样本的预处理前后对比效果如图2所示。

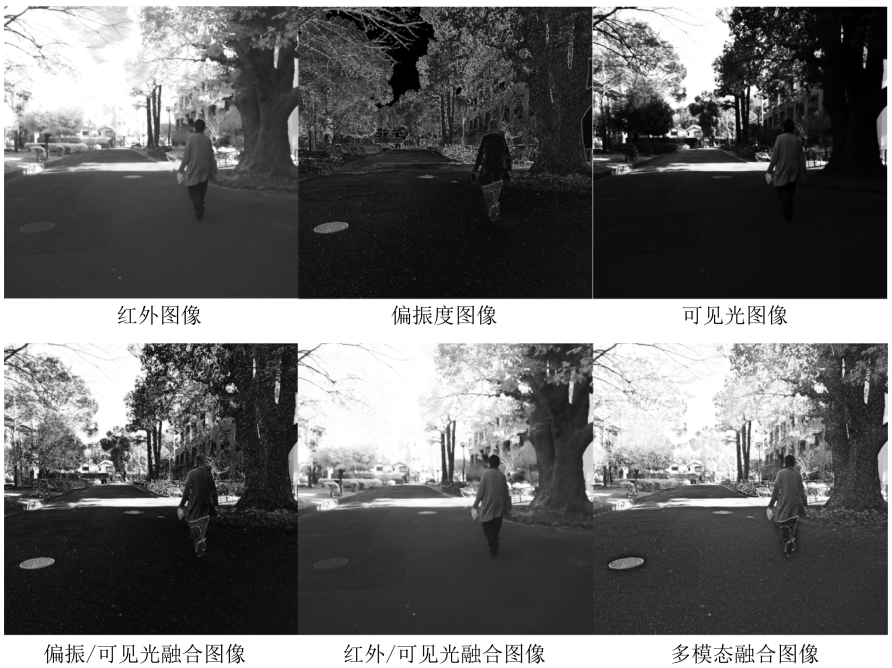



**图2 红外、偏振度以及可见光图像裁剪配准前后对比图**

**Fig. 2 Infrared, polarization degree and visible images: before and after cropping & registration**

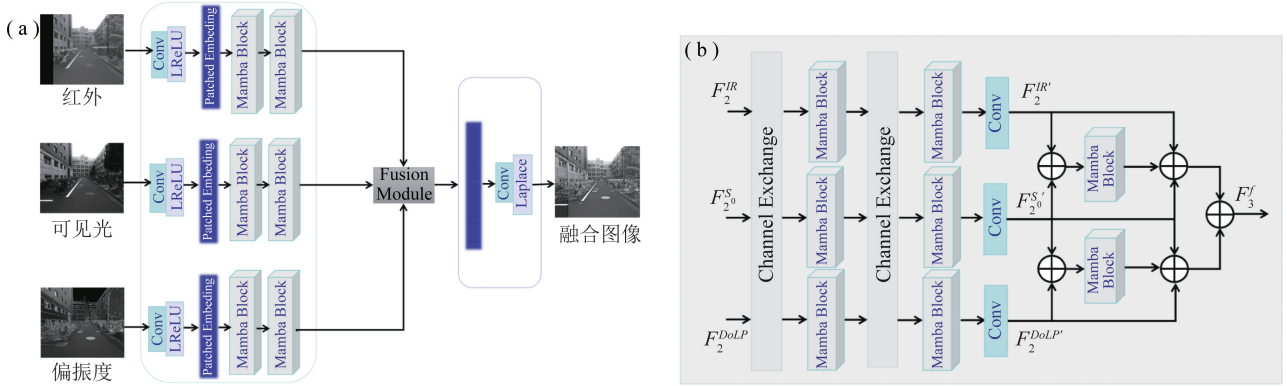
## 3 多模态融合网络架构设计

### 3.1 网络总体架构

本文构建的三模态图像融合网络总体架构如图3(a)所示,主要由多分支特征提取模块、多模态特征融合模块以及图像重建模块三个核心部分协同构成。该架构采用并行处理流的形式,旨在深度挖掘并有效耦合可见光、红外与偏振模态在复杂低光环境下的异构互补特征。

针对低光场景下信噪比极低、有效光强严重匮乏的挑战,本文在特征提取阶段采用了三路独立并行的分支网络拓扑结构。该设计可有效避免不同物理机理的模态在浅层特征空间中,因数据分布差异过大而引发的特征混叠与噪声级联放大问题。以红外特征提取分支为例,红外成像主要记录场景目标的热辐射分布,其在复杂背景及杂散光干扰下具备极高的对比度与鲁棒性,但纹理细节相对匮乏。因此,输入图像首先经由卷积层完成浅层空间特征映射,随后馈入Mamba模块进行全局长程依赖建模,最终提取出具有显著热辐射表征的深层红外特征。该独立分支在物理表征上保持了高度的封闭性,有效抑制了其他模态杂乱背景特征的干扰,确保了热目标显著信息的纯粹性。

在特征提取阶段，源图像首先输入至卷积核大小为 3×3 且步长为 1 的卷积层，并配合 Leaky ReLU 激活函数，提取图像的局部空间特征$\boldsymbol{F}_1$。随后，对$\boldsymbol{F}_1$ 执行块嵌入（Patch Embedding）操作以序列化空间维度，并将其送入 Mamba 块实现高效的高维序列特征建模。本文三条并行分支均以 Mamba 模块作为核心建模单元。如图 1 所示，输入特征经层归一化（Layer Normalization）处理后，使用两个多层感知机在两个分支上分别将特征映射为 $a$ 和 $b$。在第一个分支上，$a$ 经过卷积操作和 SiLU 激活函数处理后得到 $a'$，然后通过状态空间模型计算得到 $h$。在另一分支上，$b$ 经过 SiLU 激活函数处理后作为门控因子对 $h$ 进行门控操作得到 $h'$。最后经过多层感知机处理，并且引入残差连接，完成高级特征$\boldsymbol{F}_2$ 的提取。




注：(a) 整体网络架构；(b) 融合模块。

**图 3　多模态图像融合网络架构**

**Fig. 3　Multimodal image fusion network architecture**

为确保在多模态融合过程中保留可见光图像的纹理细节，本文在设计融合策略时，为避免引入额外参数和计算负担，引入了一种无参数的通道交换（Channel Exchange）机制，在不增加模型复杂度的前提下实现模态间的信息互通[30]，确立以可见光特征为核心、双路交互并行的思路。信息论层面，通道交换通过无参数的特征置换，促进可见光、红外与偏振三个分支的特征分布对齐，实现了隐式的域自适应，有效缩小了不同模态间的特征分布差异；物理层面，以可见光纹理为信息枢纽，将红外的热目标显著性与偏振的边缘/材质信息高效注入可见光分支，符合“可见光提供结构基底、红外提供对比度、偏振提供细节”的物理互补逻辑；设计原则层面，借鉴变化检测领域 Changer 架构的经验，本文在解码器阶段采用通道交换（此时特征语义抽象程度高，适合通道级交互），并延续非学习式硬交换策略以保持梯度独立传播、避免梯度混淆与过拟合。具体融合流程如图 3(b)所示。考虑到可见光图像在结构表达上的基石作用，三路分支特征并非简单叠加，而是以可见光特征为纽带，分别进行红外/可见光与偏振度/可见光的两两交互融合。这种设计使网络在保留红外热目标与偏振物理特性的同时，最大限度地强化图像的纹理边缘。最终，通过将上述两路交互后的融合特征进行加权求和，实现三模态信息的深度协同建模。这种层次化融合策略不仅有效解决了多模态间的信息冗余冲突，也显著提升了模型在极低照度环境中的鲁棒性。

最后，经过重建模块非线性映射与拉普拉斯增强层优化后，系统输出最终的三模态融合图像。该图像在视觉表现上不仅保留了红外模态的热目标显著性，还兼具了可见光的高频纹理与偏振模态的表面物理特征。这种端到端的重建设计确保了模型生成的图像在复杂低光环境下依然能够保持清晰的几何轮廓与丰富的语义信息。

### 3.2　损失函数设计

为了引导网络在训练过程中兼顾结构一致性、纹理细节与偏振物理特性的表达，针对偏振度与可见光融合任务，本文构建了一种多准则复合损失函数。该损失函数通过在像素、梯度及结构等多个维度对融合图像进行约束，确保生成结果在保留源图像背景信息的同时，能够有效突出偏振目标的显著性。损失函数由以下四个部分协同构成。

（1）结构一致性损失（Structural Similarity Index Measure Loss，SSIM Loss）：在图像融合任务中，保持源图像的整体拓扑结构是避免几何失真与结构畸变的关键。引入 SSIM 损失指标 $\rho_{\mathrm{SSIM}}$ 来度量融合图像与

源图像之间的结构关联。考虑到偏振度图像与可见光图像具有不同的结构特征,结构一致性损失定义为

$$\mathcal{L}_{\text{SSIM}}=\varepsilon_1(1-\rho_{\text{SSIM}}(I_{\text{F}},I_{\text{DoLP}}))+\varepsilon_2(1-\rho_{\text{SSIM}}(I_{\text{F}},I_{S_0}))\ , \tag{3}$$

式中 $I_{\text{F}}$ 表示偏振度与可见光融合图像,$I_{\text{DoLP}}$ 与 $I_{S_0}$ 分别表示偏振度图像与可见光图像;$\varepsilon_1$ 和 $\varepsilon_2$ 为权重系数,用于平衡不同模态对结构约束的影响。该损失项可保证融合结果在亮度分布和结构轮廓上与源图像保持一致。

(2) 纹理保持损失(Texture Loss):融合过程中,高频纹理信息(如边缘、细节结构)容易受到平滑操作的影响而被削弱。为增强融合图像中的细节表达能力,本文引入基于梯度的纹理约束,通过对图像梯度幅值进行约束以突出显著纹理特征。纹理保持损失 $\mathcal{L}_{\text{text}}$ 定义为

$$\mathcal{L}_{\text{text}}=\|\ \nabla I_{\text{F}}-\max(|\nabla I_{\text{DoLP}}|,|\nabla I_{S_0}|)\|_1\ , \tag{4}$$

式中$\nabla$表示 Sobel 算子计算得到的梯度信息,max(·)操作用于从不同模态中选取响应最强的纹理特征。通过该约束,融合图像能够更充分地保留源图像中显著的边缘与细节信息。

(3) 偏振特征保持损失(Polarization Feature Preservation Loss):偏振信息在目标判别中具有独特优势,尤其是偏振图像特有的局部对比度。为了防止偏振特征在融合过程中被强度信息稀释,本文设计了一种多尺度约束机制

$$\begin{aligned}\mathcal{L}_{\text{Pol}}=&\delta_1\|\ \nabla I_{\text{F}}-\max(\nabla I_{\text{F}},\nabla I_{\text{DoLP}})\|_1+\delta_2\|\Delta I_{\text{F}}-|\ \Delta I_{\text{DoLP}}\ |\ \|_1\\&+\delta_3\|\sigma_{\text{Patch}}(I_{\text{F}})-\max(\sigma_{\text{Patch}}(I_{\text{F}}),\sigma_{\text{Patch}}(I_{\text{DoLP}}))\|_1,\end{aligned} \tag{5}$$

式中 $\Delta$ 表示离散拉普拉斯算子,用于提取二阶导数信息;$\sigma_{\text{Patch}}$ 表示局部区域的标准差,用于衡量局部对比度;$\delta_1$、$\delta_2$ 与 $\delta_3$ 为权重系数。

(4) 亮度保真损失(Intensity Fidelity Loss):弱光场景常伴有光照不均或极低照度的情况,融合结果容易出现亮度偏移或失真。为了使融合图像的辐射特性更加合理,本文引入了带有区域掩模(Mask)的亮度保真损失

$$\mathcal{L}_{\text{intensity}}=\omega_1\|I_{\text{F}}-I_{S_0}\|_1+\omega_2\|(I_{\text{F}}-I_{S_0})\odot M_{\text{bright}}\|_1+\omega_3\|I_{\text{F}}-(\varphi_1 I_{S_0}+\varphi_2 I_{\text{DoLP}})\odot M_{\text{dark}}\|_1, \tag{6}$$

式中 $M_{\text{bright}}$ 与 $M_{\text{dark}}$ 分别表示通过阈值分割得到的亮度区域与低照度区域掩模,$\odot$ 为哈达玛积(Hadamard Product),$\omega_i$ 与 $\varphi_j$ ($i,j=1,2$)为权重参数。通过该损失项,网络能够对不同亮度区域采用差异化的约束权重,有效提升了复杂光照条件下的视觉一致性。

综合上述各项约束,偏振度与可见光图像融合网络训练阶段的总损失函数定义为

$$\mathcal{L}_{\text{pol-}S_0}=\gamma_1\mathcal{L}_{\text{SSIM}}+\gamma_2\mathcal{L}_{\text{text}}+\gamma_3\mathcal{L}_{\text{Pol}}+\gamma_4\mathcal{L}_{\text{intensity}}\ , \tag{7}$$

式中 $\gamma_i$ ($i=1,2,\cdots,4$)为调节各损失项权重的超参数。通过多目标的协同优化,模型能够在保证结构稳定的基础上,充分融合偏振特征与纹理细节,最终输出亮度合理的融合图像。

为了在保证结构一致性的基础上,进一步强化系统对目标区域的感知能力,针对三模态网络中的红外与可见光路径设计了专门的融合损失函数 $\mathcal{L}_{\text{IR-}S_0}$

$$\mathcal{L}_{\text{IR-}S_0}=\partial_1\mathcal{L}_{\text{SSIM}}+\partial_2\mathcal{L}_{\text{text}}+\partial_3\mathcal{L}_{\text{intensity}}\ , \tag{8}$$

式中 $\partial_i(i=1,2,3)$ 为权重系数。通过调节各损失项权重,可以实现不同约束之间的平衡。

结构一致性损失定义为

$$\mathcal{L}_{\text{SSIM}}=\varepsilon_1(1-\text{SSIM}(I_{\text{fusion}},I_{\text{IR}}))+\varepsilon_2(1-\text{SSIM}(I_{\text{fusion}},I_{S_0}))\ 。 \tag{9}$$

梯度损失定义为

$$\mathcal{L}_{\text{text}}=\|\ \nabla I_{\text{fusion}}-\max(|\nabla I_{\text{IR}}|,|\nabla I_{S_0}|)\|_1\ , \tag{10}$$

式中 $I_{\text{fusion}}$ 表示红外与可见光融合图像,$I_{\text{IR}}$ 表示红外图像,$|\cdot|$ 表示绝对值,$\nabla$表示用于计算图像梯度信息的 Sobel 算子,梯度损失引导网络保留红外与可见光图像中梯度的最大值。

红外图像主要反映目标的热辐射特性,在低照度环境中具有较高的稳定性。为保证融合图像具备合理的辐射亮度分布并突出热目标,本文引入亮度保真损失如下

$$\mathcal{L}_{\text{int}}=\frac{1}{HW}(\|I_{S_0}-I_{\text{fusion}}\|_1+\beta\mathbb{I}(I_{S_0}\leqslant I_{\text{IR}})\|I_{\text{IR}}-I_{\text{fusion}}\|_{\text{F}}^{2})\ , \tag{11}$$

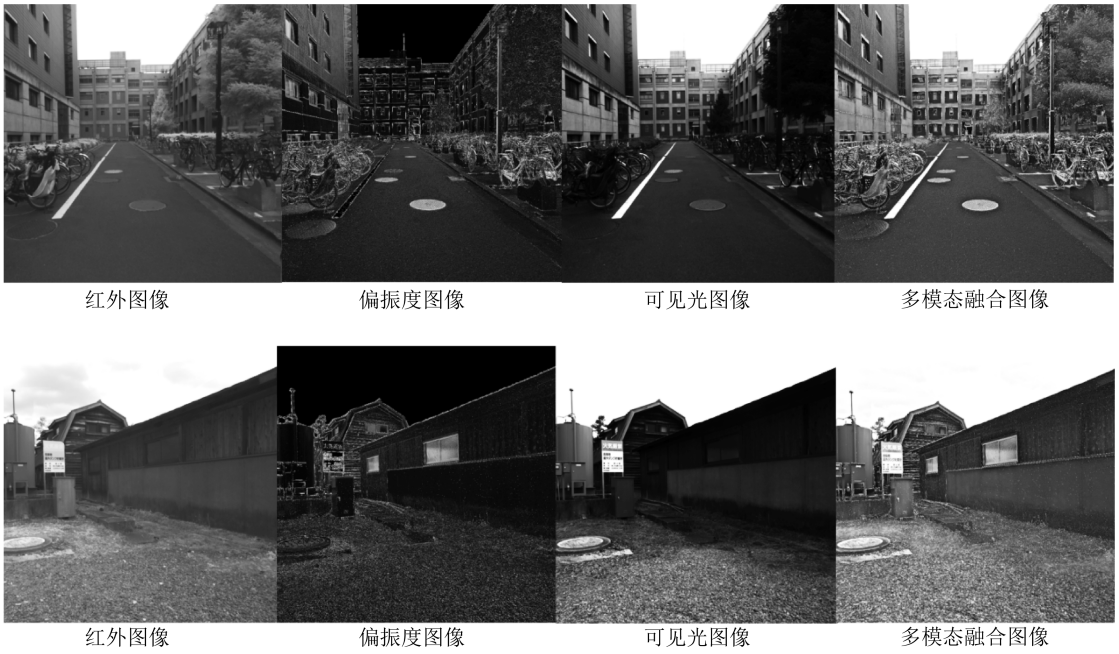
式中 $H$ 和 $W$ 分别表示图像高度和宽度，$\| I_{S_0} - I_{\text{fusion}} \|_1$ 实现背景区域保持可见光的自然感。引入条件指示函数 $\mathbb{I}(\cdot)$ 实现动态约束；当红外像素强度大于可见光（即存在热目标）时，强制融合图像像素值向红外模态趋近，从而显著提升目标的突出程度。$\beta$ 为平衡红外目标权重的系数，$\| \cdot \|_F$ 表示 Frobenius 范数。

# 4　实验结果

## 4.1　定性结果分析

为了验证本文所提三模态融合算法在低光复杂环境下的视觉感知性能，开展了多模态对比实验，结果如图 4 所示，单一模态在复杂低光环境中均存在明显表征瓶颈，而本文提出的方法在场景高保真恢复方面展现出显著优势：图 4(a)中，红外图像凭借目标与背景的热辐射差异，能够有效凸显场景中的显著目标，具备极强的环境自适应能力。然而，受限于成像机理，该模态空间分辨率较低，对比度的缺失导致场景层次感弱化，难以精细刻画复杂的几何结构。偏振度图像则对物体表面材质及粗糙度差异敏感。对于具有强方向性反射特性的结构（如墙面缝隙、井盖及金属框架），偏振模态提供了极具锐度的边缘特征，有助于实现材质区分。但其动态范围较窄，在天空或阴影等背景区域存在严重的信息缺失现象，导致场景语义呈现断层。可见光图像虽具备最高的分辨率与最符合人类视觉感知的纹理特征，但在极端低照度或阴影遮挡下，图像信号信噪比骤降，导致自行车密集区域等暗部细节被高频噪声彻底淹没，其可靠性大幅衰减。

相比之下，本文方法实现了三模态信息的深度互补：一方面，融合图像在保留可见光高频细节的基础上，结合偏振模态的边缘增强作用，使道路标线更锐利、物体几何结构更清晰，有效弥补了红外图像边缘模糊的固有缺陷；另一方面，引入红外热辐射特征，在极弱光照区域显著提升了目标对比度，消除了阴影掩蔽干扰，使得隐匿于暗部的远距离目标清晰可见，大幅增强了低光场景下的目标显著性。




注：(a) 场景 1；(b) 场景 2。

**图 4　场景 1,2 下红外、偏振度、可见光以及多模态融合图像对比图**

**Fig. 4　Comparison of infrared images, polarization degree images, visible images and multimodal fused images in Scene 1 and Scene 2**

通过对比图 4(b)中的指示牌区域，可更直观地观察到各模态在微观特征刻画上的差异：红外图像中指示牌整体轮廓模糊，受制于空间带宽限制，牌面字符等高频细节完全丢失。偏振度图像得益于偏振信息对边缘轮廓的敏感性，准确重构了指示牌边界与字符外形。可见光图像则在光照允许的范围内保留了牌面材质纹理。最终的融合结果不仅精准继承了偏振的几何轮廓与可见光的表面纹理，还借助红外信息提升了整体对比度，充分证明了三模态协同感知在重构复杂场景方面的工程价值。

## 4.2　定量结果分析

为了客观评价多模态融合算法的性能，本文选取了信息熵（Entropy，EN）、标准差（Standard Deviation，

SD)、空间频率(Spatial Frequency, SF)以及平均梯度(Average Gradient, AG)四项指标[31],对不同场景下的融合结果进行了定量分析,如表 1 所示。

从客观评估结果可知,三模态融合图像相较于单一模态图像展现出显著优势。在信息丰富度方面,本文融合图像的 EN 指标在两组场景中均达到最大值,证明通过多尺度引入红外与偏振特征,图像包含的信息量得到了最大化的提升。在细节表现力上,三模态融合结果的 SF 与 AG 指标表现优异,充分验证了融合机制对局部纹理梯度与高频边缘的增强效能。

值得注意的是,可见光图像在 SD(标准差)指标上虽表现突出,但这主要归因于可见光模态在低照度环境下较高的背景噪声或局部光强对比。相比之下,三模态融合图像的 SD 值处于次优水平且表现稳健,说明模型在有效整合各模态信息的同时,较好地抑制了极端光照带来的负面影响。而偏振度图像的 AG 略高于融合图像,是因为偏振图像在特定金属或强反射物体边缘会有极高的响应(陡峭的边缘),而融合图像为了兼顾红外的热分布和可见光的平滑纹理,会对这种极端边缘进行一定的权重平衡。

综合各项指标分析,本文提出的多模态融合方法能实现对场景信息的透彻感知,其性能表现与前述主观评价结果保持高度一致。

**表 1 场景 1、2 的定量评价结果**

**Tab. 1 Quantitative evaluation results of scene 1 and scene 2**

| 场景 | 模态 | EN | SD | SF | AG |
|---|---|---|---|---|---|
| 场景 1 | 红外图像 | 6.302 3 | 56.144 0 | 5.134 7 | 0.016 8 |
| | 偏振度图像 | 6.604 2 | 37.479 7 | **12.066 9** | **0.112 7** |
| | 可见光图像 | 6.531 7 | **69.637 5** | 5.418 3 | 0.020 6 |
| | 多模态融合图像 | **6.947 4** | 65.000 2 | 11.695 9 | 0.111 0 |
| 场景 2 | 红外图像 | 6.663 9 | 72.665 3 | 6.212 3 | 0.020 2 |
| | 偏振度图像 | 5.920 8 | 36.563 9 | **11.348 5** | **0.108 9** |
| | 可见光图像 | 6.283 7 | **81.623 7** | 7.231 3 | 0.030 4 |
| | 多模态融合图像 | **6.691 7** | 79.281 2 | 10.833 0 | 0.090 6 |

### 4.3 消融实验分析

为验证网络架构中各模态引入的必要性,开展了消融对比实验。图 5 展示了场景 3 的原始源图像及不同融合策略的可视化结果。由单模态成像可知:红外图像在远处树木等背景区域出现明显的过饱和失真,导致高光区域辐射细节丢失;偏振度图像成功剥离了材质轮廓,但整体光度分布极不均衡;可见光图像在极低光照下出现信噪比急剧下降,纹理特征近乎失效。

针对上述单模态成像存在的固有缺陷,本文对比了偏振度-可见光双模态、红外-可见光双模态与本文三模态架构的融合表现。由图 5 可见,不同融合策略在行人显著性及场景上下文解算上差异极大:仅融合偏振度与可见光,只能勾勒行人粗略轮廓,暗部衣物褶皱等关键纹理缺失,目标与背景高保真解耦失败;仅融合红外与可见光,虽突出了热目标,但因缺乏偏振矢量特征引导,井盖及金属隔离栏等非热源显著物体辨识度低下;而本文的红外-偏振度-可见光三模态网络,以红外热辐射锚定全局显著性分布,以偏振高频响应锐化边缘拓扑,并以可见光纹理进行底层基底映射。红外模态在此并非进行简单的像素强加,而是在目标层级实现了全局对比度重塑,彻底弥补了无红外方案在低光场景下的感知盲区。

为量化模态增益,表 2 列出了场景 3 的消融客观指标。在信息承载力方面,三模态图像的 EN 值达到 7.015 5,较偏振度单模态与偏振-可见双模态分别提升 5.18%与 0.67%,确证了热辐射信息的引入有效打破了原有的信息承载上限。从对比度重构来看,红外单模态因背景过曝引发剧烈极化,其 SD 值代表的是畸变虚高而非有效对比度;三模态融合依托亮度保真约束(Intensity Fidelity),在规避失真的前提下,使 SD 值跃升至 78.270 3,显著优于偏振-可见光双模态的 75.544 5。在细节特征保留上,三模态图像的 SF 与 AG 指标紧随偏振单模态,远超红外-可见光双模态,证明网络在大幅拉升全局对比度的同时,通过通道交互机制近

乎无损地保留了偏振诱导的微观形貌特征。消融实验结果验证了三模态协同建模策略在低光复杂环境探测中的有效性。

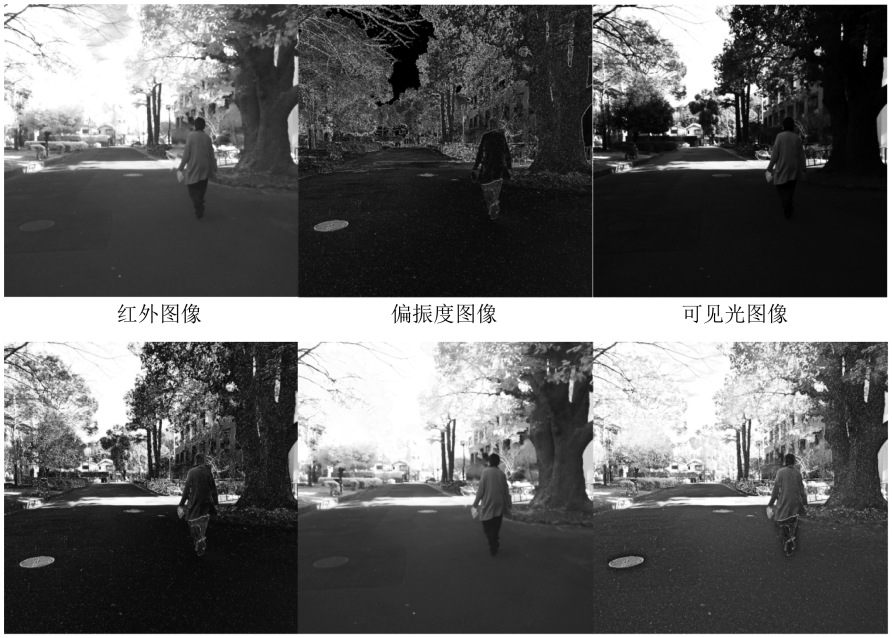



**图 5　场景 3 的红外、偏振度、可见光图像原始图和消融对比图**

**Fig. 5　Original images and ablation comparison images of infrared, polarization degree and visible images under Scene 3**

**表 2　场景 3 的定量评价结果**

**Tab. 2　Quantitative evaluation results for scene 3**

| 类别 | EN | SD | SF | AG |
|---|---|---|---|---|
| 红外图像 | 6.287 4 | 76.674 1 | 6.155 4 | 0.025 4 |
| 偏振度图像 | 6.670 5 | 42.251 6 | **12.719 5** | **0.152 1** |
| 可见光图像 | 6.366 7 | 71.480 5 | 6.725 6 | 0.034 3 |
| 偏振度-可见光融合 | 6.968 8 | 75.544 5 | 12.522 3 | 0.132 4 |
| 红外-可见光融合 | 6.458 2 | **81.386 8** | 6.407 1 | 0.026 8 |
| 多模态融合图像 | **7.015 5** | 78.270 3 | 12.122 0 | 0.116 8 |

## 5　结论

针对复杂低光环境下单一模态视觉感知受限的问题，本文提出了一种可见光-红外-偏振三模态协同感知与融合网络。在特征提取端，构建了多分支并行网络架构，并引入 Mamba 状态空间模型，有效实现了多源异构特征的全局长程依赖建模；在特征融合端，确立了以可见光纹理为锚点的双路融合策略；在优化约束端，设计多准则复合损失函数，引导模型在特征增强与噪声抑制之间实现最优解。定性与定量实验结果全面验证了该架构的有效性与鲁棒性。定量评价表明，三模态融合图像在信息熵（EN）上达到全类别最优，较偏振度单模态及偏振度-可见光双模态分别提升 5.18%与 0.67%，突破了单一或双模态的信息承载上限；同时，在空间频率（SF）与平均梯度（AG）上维持了极具竞争力的高频特征保真度。消融实验进一步确证，红外热辐射特征的引入并非简单的通道叠加，而是重塑暗区目标显著性与全局对比度（SD）的关键。综上所述，本文构建的三模态融合框架有效打破了极端光照与杂散光干扰的限制，为后续的目标识别与检测、视觉导航等下游任务提供了高质量的数据支撑。

## 参 考 文 献


[1] Zhang X, Wang X, Yan C D, et al. EV-fusion: a novel infrared and low-light color visible image fusion network integrating unsupervised visible image enhancement[J]. IEEE Sensors Journal, 2024, 24(4): 4920-4934.

[2] 丁政泽, 聂仁灿, 李锦涛, 等. MTFuse:基于 Mamba 和 Transformer 的红外与可见光图像融合网络[J]. 计算机科学, 2025, 52(8): 188-194.

[3] 沈英, 黄春红, 黄峰,等. 红外与可见光图像融合技术的研究进展[J]. 红外与激光工程, 2021, 50(9): 152-169.

[4] Yang Y Q, Ran W H, Li Y, et al. Multi-dimensional visual information processing under complex light environments using time-evolved polarization-sensitive synaptic electronics[J]. Nature Communications, 2025, 16(1): 5665.

[5] Li D, Ma D, Guo K, et al. Polarization characteristics motivating target detection in different polarization spaces[J]. Optics & Laser Technology, 2024, 171: 110430.

[6] Zhong B B, Wang X, Wang D Q, et al. Target-background contrast enhancement based on a multi-channel polarization distance model [J]. Bioinspiration & Biomimetics, 2021, 16 (4):046009.

[7] Xiang K, Yang K, Wang K. Polarization-driven semantic segmentation via efficient attention-bridged fusion[J]. Optics Express, 2021, 29(4): 4802-4820.

[8] He C, He H H, Chang J T, et al. Polarisation optics for biomedical and clinical applications: a review[J]. Light: Science & Applications, 2021, 10: 194.

[9] Duan J, Zhang H, Liu J, et al. A dual-weighted polarization image fusion method based on quality assessment and attention mechanisms [J]. Frontiers in Physics, 2023, 11: 1214206.

[10] Karim S, Tong G, Li J Y, et al. Current advances and future perspectives of image fusion: a comprehensive review[J]. Information Fusion, 2023, 90: 185-217.

[11] Li G F, Lin Y J, Qu X D. An infrared and visible image fusion method based on multi-scale transformation and norm optimization[J]. Information Fusion, 2021, 71: 109-129.

[12] Hu M Q, Sun B, Kang X D, et al. Multiscale structural feature transform for multi-modal image matching[J]. Information Fusion, 2023, 95: 341-354.

[13] Tang D, Xiong Q Y, Yin H P, et al. A novel sparse representation based fusion approach for multi-focus images[J]. Expert Systems with Applications, 2022, 197: 116737.

[14] 吴强, 王新赛, 王卫平, 等. 基于 LLE 的红外成像彩色化处理算法研究[J]. 应用光学, 2012, 33(4): 733-737.

[15] Liang P W, Jiang J J, Liu X M, et al. Fusion from decomposition: a self-supervised decomposition approach for image fusion[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 719-735.

[16] Liang Y P, Wakaki R, Nobuhara S, et al. Multimodal material segmentation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 19768-19776.

[17] Zhao Z, Xu S, Zhang C, et al. DIDFuse: Deep image decomposition for infrared and visible image fusion[C]//Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (IJCAI), Virtual Event. 2020, 970-976.

[18] Jian L H, Yang X M, Liu Z, et al. SEDRFuse: a symmetric encoder-decoder with residual block network for infrared and visible image fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-15.

[19] Zhang J C, Shao J B, Chen J L, et al. PFNet: an unsupervised deep network for polarization image fusion[J]. Optics Letters, 2020, 45 (6): 1507-1510.

[20] Zhang Y, Liu Y, Sun P, et al. IFCNN: a general image fusion framework based on convolutional neural network[J]. Information Fusion, 2020, 54: 99-118.

[21] Zhang X C, Demiris Y. Visible and infrared image fusion using deep learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 10535-10554.

[22] Safari M, Fatemi A, Archambault L. MedFusionGAN: multimodal medical image fusion using an unsupervised deep generative adversarial network[J]. BMC Medical Imaging, 2023, 23(1): 203.

[23] Li K Y, Qi M B, Zhuang S, et al. TIPFNet: a transformer-based infrared polarization image fusion network[J]. Optics Letters, 2022, 47(16): 4255-4258.

[24] Qu L H, Liu S I, Wang M N, et al. TransMEF: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning[C]//Proceedings of the AAAI conference on artificial intelligence. Palo Alto, USA: AAAI Press, 2022: 2126-2134.

[25] Ma J Y, Tang L F, Fan F, et al. SwinFusion: cross-domain long-range learning for general image fusion via swin transformer[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(7): 1200-1217.


(下转第 865 页)

[16] Xu B, Wang P, He Y J, et al. Leveraging structural information to improve point line visual-inertial odometry[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 3483-90.

[17] Qin T, Li P L, Shen S J. VINS-mono: a robust and versatile monocular visual-inertial state estimator[J]. IEEE Transactions on Robotics, 2018, 34(4): 1004-20.

[18] Shi J. Good features to track[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1994, 593-600.

[19] Tomasi C. Detection and tracking of point features[J]. Technical Report, 1991, 91(21): 9795-9802.

[20] Zhang L L, Koch R. An efficient and robust line segment matching approach based on LBD descriptor and pairwise geometric consistency [J]. Journal of Visual Communication and Image Representation, 2013, 24(7): 794-805.

[21] 张彧，檀祖冰，曹东璞，等. 基于视觉和惯性测量单元的里程计关键技术研究综述[J]. 机械工程学报，2024，60(10)：3-21.

[22] Bartoli A, Sturm P. Structure-from-motion using lines: representation, triangulation, and bundle adjustment[J]. Computer Vision and Image Understanding, 2005, 100(3): 416-441.

[23] Si H Q, Yu H, Chen K Y, et al. Point-line visual-inertial odometry with optimized line feature processing[J]. IEEE Transactions on Instrumentation and Measurement, 2024:73:1-13.

[24] Burri M, Nikolic J, Gohl P, et al. The EuRoC micro aerial vehicle datasets[J]. The International Journal of Robotics Research, 2016, 35(10): 1157-1163.

[25] Lu X H, Yaoy J, Li H A, et al. 2-line exhaustive searching for real-time vanishing point estimation in Manhattan world[C]//Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2017:345-353.

[26] He Y J, Zhao J, Guo Y, et al. PL-VIO: tightly-coupled monocular visual-inertial odometry using point and line features[J]. Sensors, 2018, 18(4): 1159.

[27] Bustos A P, Chin T J, Eriksson A, et al. Visual SLAM: why bundle adjust? [C]//Proceedings of International Conference on Robotics and Automation IEEE, 2019:2385-2391.

[28] Schubert D, Goll T, Demmel N, et al. The TUM VI benchmark for evaluating visual-inertial odometry[C]//Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2019:1680-1687.

[29] Grupp M. Evo: Python package for the evaluation of odometry and slam [Z], 2017.

[30] Fu Q, Wang J L, Yu H, et al. PL-VINS: real-time monocular visual-inertial SLAM with point and line features[EB/OL]. arXiv preprint arXiv:200907462, 2020.

（责任编辑：刘庆松）

（上接第 832 页）

[26] Gu A, Johnson I, Goel K, et al. Combining recurrent, convolutional, and continuous-time models with linear state space layers[C]. Proceedings of Neural Information Processing Systems, Virtual, 2021, 34: 572—585.

[27] Smith J T H, Warrington A, Linderman S W. Simplified state space layers for sequence modeling [EB/OL]. arXiv preprint arXiv:2208.04933, 2022.

[28] Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces [EB/OL]. arXiv preprint arXiv:2111.00396, 2021.

[29] Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces [EB/OL]. arXiv preprint arXiv:2312.00752, 2023.

[30] Fang S, Li K Y, Li Z. Changer: feature interaction is what you need for change detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-11.

[31] Sun B, Gao Y X, et al. Analysis of quality objective assessment metrics for visible and infrared image fusion[J]. Journal of Image and Graphics, 2023,28(1):144-155.

（责任编辑：刘庆松）