

基于模态度量的多模态海上船舶识别方法

徐政伟^{1a,b}, 黄沛基^{1a,b}, 何建樑^{1a,b}, 王世豪²

(1.河南师范大学 a.计算机与信息工程学院(人工智能学院);b.河南省教育人工智能与个性化学习重点实验室,
河南 新乡 453007;2.哈尔滨工程大学 信息与通信工程学院,哈尔滨 150006)

摘要:现有的海上多模态识别方法在处理复杂非线性特征方面存在欠缺之处,无法充分利用不同模态的信息,而且融合过程采用的是固定融合策略、融合权重.基于此提出了一种基于模态度量的海上船舶识别网络,这个模型引进了特征过滤模块,能够抑制无关信息,减轻模态间的分布差异,增强多模态特征提取的鲁棒性.本文还设计了模态度量评估机制,依据各模态特征的贡献来动态分配融合权重,从而最大程度地减少低价值模态对最终融合性能的影响.此机制能够根据当前场景、环境的变化,自适应地调整融合权重,提高融合效果.本文提出的方法在公开的 VAIS 海上红外可见光多模态数据集上进行了基准测试,在准确性和计算效率方面都优于当前最先进的方法.

关键词:多模态融合;海上船舶识别;特征过滤;模态度量

中图分类号:TP391

文献标志码:A

文章编号:1000-2367(2026)05-0105-07

多传感器的普及应用提高了海上船舶的识别效率,从而可以对海上船舶的识别提供不同维度的信息.近年来设计可靠的海上多模态融合框架引起了工业界和学术界的广泛关注.全球各国都将海洋领域视为重要的战略发展领域,海洋目标识别技术在军民两用方面具有广阔的应用前景.船舶作为海上交通、渔业捕捞、海洋资源开发和军事侦察等多个领域的关键载体,其识别任务具有重要的研究价值.红外和可见光图像的融合在军事和民用领域有许多应用,例如视频监控、物体识别、跟踪和遥感^[1-2].一些现有的多模态融合框架往往采用静态的融合,忽略了模态之间信息的不均衡,例如可见光图像能够提供目标清晰的纹理信息,但是,当光照及其他环境因素影响时,红外图像中的热辐射信息可以克服这些限制,为本文提供更加稳定和可靠的目标信息.因此如何挖掘模态之间的互补信息越来越受到研究者的关注.在本文中重点研究用于度量不同模态贡献度的多模态融合网络.

在一般的情况下,海上船舶识别会碰到复杂的气象情况、动态的海面情况,这使目标观测、跟踪的难度不断加大,海面环境有着很大程度的不确定性,诸如波浪、光线反射、雨雾干扰等相关情况,从而造成目标信号易被噪声所淹没,并且目标的运动模式一般具有不规则特性,这给识别系统提取稳定特征带来了更大难度.传感器所接收的回波信号受海洋背景影响比较大,尤其是处于多路径效应、目标被遮挡的状况之中,精确识别面临许多难题,近年的研究显示多模态数据融合可以提高识别的准确性,不同模态数据间的不统一情况、机器同步处理的复杂性依旧是突出难题,一般的海上目标识别方法大多依赖单模态信息,也就是只使用单一类型的传感器数据对目标进行检测与分类.

收稿日期:2025-06-25;**修回日期:**2025-08-05.

基金项目:国家自然科学基金(62271499);河南省科技攻关计划项目(252102211049).

作者简介(通信作者):徐政伟(1994—),男,河南新乡人,河南师范大学讲师,博士,研究方向为多源信息融合、领域大模型,E-mail: xuzhengwei@htu.edu.cn.

引用本文:徐政伟,黄沛基,何建樑,等.基于模态度量的多模态海上船舶识别方法[J].河南师范大学学报(自然科学版),2026,54(5):105-111.(Xu Zhengwei, Huang Peiji, He Jianliang, et al. Multimodal maritime ship recognition method based on modality metric[J]. Journal of Henan Normal University(Natural Science Edition), 2026, 54(5):105-111.DOI:10.16366/j.cnki.1000-2367.2025.06.25.0001.)

1 相关工作

从红外图像特征或者可见光图像特征当中能提取特定船舶的识别特征,前者说的是船舶进行热成像时所产生的红外信号,这些信号体现了船舶热辐射方面的特性,包含了发动机所产生的热量、船体材料的温度分布等,而可见光信号特征跟船舶在自然光或者人工照明之时的反射特性相关联,含有船舶的外观轮廓、颜色、标识、结构特征等情况.传统的船舶识别方法都是基于单模态的视觉方法并结合专家知识,受到人为主观差异的影响比较大.Dong 等^[3]通过引入多尺度对比学习网络(MSCL-Net)来提取同类船舶的相似特征,并且利用区域裁剪和放大模块克服背景复杂性,以此提高海上船舶的精细分类精度.Xiao 等^[4]提出了一种渐进数据增强方法,主要致力于处理光学遥感图像数据.借助船舶仿真样本与神经风格迁移(NST)网络相结合,生成了数量众多的高质量遥感船舶图像.此方法是为解决训练数据匮乏问题,进而提高船舶图像分类的精度、多样性.He 等^[5]提出一种多尺度密集网络,通过融合不同尺度卷积特征图(MS-GBCNN),提高双极化 SAR 图像的船舶分类性能.虽然通过深度学习方法进行的研究让海上目标识别性能得到了一定提高,但是基于单模态的识别方法在面对海上复杂场景时,其所达到的精度会受到一定限制,泛化能力也有不足,特别是在云、雨、雾等恶劣天气状况下.而且这些模型对于不平衡数据、复杂环境的适应能力比较有限,同时还缺少足够的可解释性.传统单模态的海上目标识别方法在实际应用中常常效果不太理想.海上环境常常会有光照变化、雾霾、海浪等复杂状况相随,这些因素会致使可见光图像的质量降低,从而对船舶识别的准确性产生影响.红外图像虽说能够在弱光或者恶劣天气状况下有效发挥作用,但其分辨率相对而言较低,并且容易遭受其他热源的干扰.以往的研究因数据获取难度较大、质量不高^[6]等因素大多局限于单模态的海上船舶识别,伴随光学遥感技术取得进步、多传感器设备得以普及,研究者把目标转向了多模态海上船舶识别.Zhang 等^[7]建立了世界上首个公开的可见光与红外船舶图像配对数据集,称作 VAIS,为多模态船舶识别领域的发展提供了数据方面的基础.Liu 等^[8]设计了一个带有一个发生器、双鉴别器的红外目标与可见光图像融合网络,运用差异学习来提取二者的共有模态特征,同时保留了红外目标的结构信息、可见光图像的纹理细节.Tlig 等^[9]提出了 3 种阶段融合策略实现可见光和红外图像模态的增效融合,该策略分别通过早期融合、中期融合、后期融合的方法提取来自两种模态的像素级别、特征级别和决策级别的对齐特征,从多个粒度上实现两个模态的信息互补.Tang 等^[10]提出了一种实用的红外与可见光图像融合网络 PSFusion,通过渐进语义注入和场景保真来提升高层视觉任务的性能.Xie 等^[11]利用了自编码器结构,在高频、低频和中频 3 个频段上提取特征,还运用可学习的注意力机制来自适应融合特征.Zhao 等^[12]整合 Transformer 和 CNN 的优势,开发出名为 CD-DFuse 的一个网络,通过相关性驱动的特征分解手段,有效提取、融合模态特定和模态共享特征.然而,这些流行的深度学习算法一般采用固定的前向传播流程,没有顾及模态相互间的差异,处于复杂场景下也许不会发挥多模态的优势.综上所述,以上方法多是从网络层面进行模态平衡,但是没有真正从衡量一个模态的贡献度方面来设计一套动态的融合机制,且在特征的高维映射上面缺乏可解释性,使得其增加了在关键工业实际应用场景部署的风险.

2 基于模态度量的海上船舶识别模型

2.1 网络结构

MM-NET 采用双流编码-解码架构(见附录图 S1),输入为配对的可见光与红外图像.编码器包含共享特征编码器和细节特征编码器,分别提取全局语义和局部细节特征.解码器通过贡献度融合模块(CRF)动态聚合多模态信息,输出分类结果.

2.2 特征过滤模块

该特征过滤模块由两部分组成:一是基于 ResNet 模块的共享特征编码器,二是基于 Dense-KAN 模块的细节特征编码器.这两个组件协同作用,共同构建了长短距离编码器.

采集到的红外与可见光图像处理之后会被送入网络模型识别,所输入的数据是具有目标对齐的红外、可

见光图像对,首先把数据送入长程特征提取器(LFE)提取浅层特征,LFE的作用是把输入的原始图像转化成多尺度、高维度特征表征,为之后的特征融合与判别给予丰富的语义信息,LFE是基于Resnet所搭建的特征提取网络,包含多个卷积层以及池化层,以渐渐抽象图像内容:

$$(I_{\text{IR}}^{\text{B}'}, I_{\text{VIS}}^{\text{B}'}) = \text{LFE}(I_{\text{IR}}', I_{\text{VIS}}'), \quad (1)$$

其中, $(I_{\text{IR}}^{\text{B}'}, I_{\text{VIS}}^{\text{B}'})$ 代表长程特征提取器提取到的浅层基础特征. 之后将浅层基础特征导入短程特征提取器,短程特征提取器由非线性学习能力上更胜一筹的Dense-KAN模块组成,目的是对红外和可见光特征中更复杂的非线性关系加以捕捉,Dense-KAN模块作为起到特征过滤作用的短程特征提取器,能够有力捕获高频特征,解决梯度消失的问题,还能增加多模态数据融合的鲁棒性和精确性,把长程特征提取器的基础特征给到短程特征提取器模块:

$$(I_{\text{IR}}^{\text{DK}'}, I_{\text{VIS}}^{\text{DK}'}) = \text{SFE}(I_{\text{IR}}^{\text{B}'}, I_{\text{VIS}}^{\text{B}'}), \quad (2)$$

其中, $(I_{\text{IR}}^{\text{DK}'}, I_{\text{VIS}}^{\text{DK}'})$ 代表经过短程特征提取器解码的高维信息,短程特征提取器内部嵌入了一个Dense KAN Block,它是由一系列的KAN层进行密集连接组成,KAN由一系列连接的Kolmogorov-Arnold层组成,每个层包含一组可学习的一维激活函数, k 层KAN可以表现为多个KAN层的嵌套,具体如下: $\text{KAN}(x) = (\Phi_{L-1} \circ \Phi_{L-2} \circ \cdots \circ \Phi_1 \circ \Phi_0)x$, 其中, Φ_i 表示整个KAN网络的第*i*层. 每个KAN层,具有 n_{in} 维输入 n_{out} 维输出, Φ 包含 $n_{\text{in}} \times n_{\text{out}}$ 可学习的激活函数 $\Phi: \Phi = \{\phi_{q,p}\}$, $p=1, 2, \dots, n_{\text{in}}, q=1, 2, \dots, n_{\text{out}}$, 其中, $\phi_{q,p}$ 是 p 到 q 之间的特征映射. 同时,为了避免网络在训练过程中关键信息的流失,设计了一种Dense-KAN模块,帮助每一层更好地利用前面层的输出,从而进一步丰富模型的特征表示能力,具体如下.

对于第 l 层,其输入是所有前面层输出的拼接: $\mathbf{H}_l = [\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{l-1}]$, 其中, $[\cdot]$ 表示将各层的输出进行拼接.

第 l 层的输出由其Kolmogorov-Arnold模块计算得出: $\mathbf{y}_l = f_l(\mathbf{H}_l)$, 其中, $f_l(\cdot)$ 代表第 l 个Kolmogorov-Arnold模块所表示的非线性变换.

最终网络的每一层进行密集连接,经过Dense-KAN模块处理后的每个模态最后的输出为: $f^m = \mathbf{y}_L = f_L([\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{L-1}])$, 其中 f^m 是Dense-KAN模块最后的输出, $m \in \{\text{IR}, \text{VIS}\}$.

2.3 模态度量融合模块

多模态数据融合的主要难题之一是怎样有效分配不同模态特征的贡献份额,传统的办法一般给各模态特征给予相等的权重,或通过简单的线性加权形式进行融合,这在处理复杂环境当中的非均衡数据时,一般会造成模型性能下降,为解决这个问题,本文研发了Contribution Fuse模块,本文所设计的contribution fuse模块和以往不太一样,以两个视角来开展:a)能量函数捕捉模态的不确定性,b)置信估计函数捕捉单模态决策的可信度.

在角度a)中,为了获得有效的不确定性度量,本文使用能量分数,这是在不确定性学习研究方向主流的度量方法.对于多模态红外可见光数据,不同模态的密度函数可以通过相应的密度函数来估计.

在多模态数据融合的过程当中,不同模态的特征其质量和可靠性可能不一样,为了好好衡量每个模态的不确定性,本文采用了基于熵相关的能量函数,熵衡量概率分布所具有的不确定性或随机性,说明系统的不可预测程度越高,基于熵所形成的能量函数如下所示: $H_{\text{IR}} = - \sum p_{\text{IR}}(X_{\text{IR}}) \lg p_{\text{IR}}(X_{\text{IR}})$.

熵函数针对每个模态来度量其“不确定性”,当熵值比较高的时候,就表明该模态的决策越具有不确定性.在实际应用场景当中,熵能够借助模型的输出概率予以估计,或者依靠训练数据来展开推断.为了实现不同

模态熵的公平比较,本文针对每个模态的熵值进行归一化处理: $\hat{H}_{\text{IR}} = \frac{H_{\text{IR}}}{H_{\text{IR}} + H_{\text{VIS}}}$, $\hat{H}_{\text{VIS}} = \frac{H_{\text{VIS}}}{H_{\text{IR}} + H_{\text{VIS}}}$.

在模态融合开展的进程当中,除了要思索不确定性之外,单模态决策所具有的可信度也会对最终的融合结果产生影响作用,置信度度量能让我们了解每个模态输出于整体决策中的影响力情况,要是捕获这一要点,可运用基于概率的置信度估计办法,设定每一种模态的预测输出是 $c_{\text{IR}} = \sigma(W_{\text{IR}}^T \mathbf{f}_{\text{IR}})$, $c_{\text{VIS}} = \sigma(W_{\text{VIS}}^T \mathbf{f}_{\text{VIS}})$, 在此之中,所涉及的 σ 是Sigmoid函数,而与之相关的 $\mathbf{f}_{\text{IR}}$ 和 $\mathbf{f}_{\text{VIS}}$ 分别为红外、可见光模态的特征向量,另外 W_{IR} 和 W_{VIS} 则是权重参数.借助这些置信度值,能够对每个模态的贡献度予以评估,进而依据置信度来进行

加权融合。

最终的模态融合系数是由熵函数、置信度估计共同决定得出的。为了能够全面考虑模态的不确定性还有置信度这两方面因素,在本文中,首先会通过熵值进行归一化处理、对置信度进行计算,从而得出每个模态的融合系数:由上述公式所计算得出的融合系数 α_{IR} 和 α_{VIS} 体现出了红外模态、可见光模态在最终融合进程中的贡献占比情况。这些系数对模态的熵(也就是不确定性的一种度量方式)、置信度(关乎决策的可信度)都予以了考量,进而保证每一个模态针对最终决策所做出的贡献是合理的并且能够进行动态调整。最终而言,融合过后的特征向量 $\mathbf{f}_{\text{fusion}}$ 能够按照如下方式予以表示: $\mathbf{f}_{\text{fusion}} = \alpha_{\text{IR}} \mathbf{C}_{\text{IR}} \mathbf{f}_{\text{IR}} + \alpha_{\text{VIS}} \mathbf{C}_{\text{VIS}} \mathbf{f}_{\text{VIS}}$ 。

Contribution Fuse 模块凭借引入熵函数以及置信度估计,得以给每个模态安排恰当的融合权重,在不确定性比较高或者置信度不高的模态当中,系统会自动减少其对融合结果所造成的作用,而若模态的可靠性比较高的话,就会增大其对结果的贡献,这样一种动态加权机制使多模态数据融合的鲁棒性和精度有所提高,特别是在复杂且不确定性相当高的环境当中。

2.4 Loss 函数

鉴于在没有额外正则化时这种模态度量估计无法给出准确估计,设计出一种基于历史轨迹信息的正则化方法,以此提高模态贡献度分数与相应损失间的相关性,设计一种借助历史轨迹信息的正则化方法,将训练阶段的样本损失当作监督信息,进而利用历史损失函数来规则化融合权重,对于给定的一个样本的 m 模态样本 (x_i, y_i) , 计算训练的平均损失: $\Psi_i^m = \frac{1}{T} \sum_{t=T}^{T_s+T} \ell(y_i, f_{\theta_t}^m(x_i))$, 其中, $f_{\theta_t}^m$ 是具有参数 θ_t 的每个迭代时期 t 上的单模态分类器。在训练 $T_s - 1$ 个 epoch 之后,采样 T 次并计算平均训练损失。并通过以下关系确定融合权重: $\Psi_i^m \geq \Psi_j^m \rightarrow w_i^m \leq w_j^m$ 。

完整正则化的定义如下: $L_{\text{reg}} = \max(0, g(w_i^m, w_j^m)(\kappa_i^m - \kappa_j^m) + |w_i^m - w_j^m|)$,

$$g(w_i^m, w_j^m) = \begin{cases} 1 \rightarrow w_i^m > w_j^m, \\ 0 \rightarrow w_i^m = w_j^m, \\ -1 \rightarrow \text{其他}. \end{cases}$$

最后,将总的损失函数定义为每个模态的交叉熵损失以及正则化项的和:

$$L_{\text{overall}} = L_{\text{CE}}(y, f(x)) + \sum_{m=1}^M L_{\text{CE}}(y, f^m(x^m)) + \lambda L_{\text{reg}},$$

其中 λ 是控制正则化强度的超参数, L_{CE} 和 L_{reg} 分别是交叉熵损失和正则化项。

3 实验结果与分析

3.1 实验数据与实验环境

在本节中,本文通过在 VAIS(一个公共海上多模态船舶数据集)上进行实验来评估所提出的 MM-NET 方法的性能。实验平台使用了 Pytorch 开源软件库和 NVIDIA RTX4060Ti。同时, Average precision、Average recall、Average F1-score、Average-precision 被用作评价分类结果的指标。数据集原始文献和获取信息参见参考文献[7]。实验平台使用 Python 3.8 编程语言,基于 PyTorch 2.0 深度学习框架搭建模型,硬件环境为 Intel Core i7-13700K 处理器、NVIDIA RTX 4060Ti 显卡。数据预处理及特征可视化采用 OpenCV 4.8.0 和 Matplotlib 3.7.1 工具库,模型训练与评估的计算加速依赖 CUDA 11.8 及 cuDNN 8.9。

3.2 评价指标与数据集介绍

VAIS^[7]数据集被看作是全球首个向公众开放的可见光与红外船舶图像配对数据集,里面有 2 865 张图像。该数据集由 1 623 张可见光图像、1 242 张红外图像共同组成,其中有 1 088 对相互对应的图像。而且,这个集合还包括 154 张日间红外图像。数据集中有 264 种不同的船只,被分成 6 个大致类别,这些类别又进一步细分为 15 个细分类别。这些类别包括非商业船只、帆船、中型客船、中型“其他”船只、拖船、小型船只。本文在数据集中建立了可见光与红外图像之间的目标级关联,让可见光和红外图像对能够精准对齐,同时去除损坏的或未标记的图像数据,并且剔除仅包括单一模态图像的目标样本。

3.3 结果分析

为验证本文方法有效,进行了广泛实验.用于比较的算法大体分为两类:一是单模态识别方法,二是最先进的多模态识别方法.单模态分类方法包括红外图像分类与纯可见光图像分类.单模态方法有基于 Transformer 的模型像 VIT(视觉 Transformer)^[13],还有基于 CNN 的模型如 ResNet^[14]、MambaOut^[15].具体来讲,VIT 是首个把 Transformer 架构用于计算机视觉任务的模型,它把图像划分成固定大小的块,将块序列输入标准 Transformer 处理来实现图像分类.Swin Transformer^[16]属于分层 Transformer 模型,其作用是在视觉任务里对多尺度特征进行更有效的处理.ResNet^[14]是深度卷积神经网络,借助残差块减轻深度网络里的梯度消失问题.MambaOut^[15]作为 CNN 架构的一种变体,重点在于高效特征提取、计算优化.最新多模态分类方法有基于动态融合的,还有处理模态不平衡的梯度调制技术,例如 OGM-GE^[17]、CAFN^[18]、QMF^[19]、Transfuser^[20]、MMTM^[21].OGM-G 属于动态梯度调制方法,它会依照不平衡的模态贡献动态调整优化的过程,从而处理多模态学习里优化的不平衡问题.CAFN 是协同注意力融合网络,通过协同注意力模块达成模态间的相互带动,还依靠动态像素级融合来生成更有细粒度的多模态特征表示.QMF 作为质量驱动多模态融合的框架,其引入动态融合机制来提升模型在低质量多模态数据情形当中的鲁棒性与泛化能力.TransFuser 是基于 Transformer 的多模态融合模型,运用自注意力机制整合图像、激光雷达的特征,以此提高复杂动态环境下的性能.MMTM 是多模态转移模块中的一种,借助挤压及激励机制,对多模态 CNN 通道特征予以动态更新的校准,从而达成高效的跨模态信息互通.

在这一块内容当中,会把本文所提的 MM-NET 和最先进的单模态、多模态识别方法来做比较,这些方法在 VAIS 数据集上开展评估工作,把精确率、召回率、F1 分数还有准确率用作评估指标,表 1 呈现出来不同单模态识别方法在海上目标识别方面的性能,从结果可以清楚地看到相同方法在不同模态下的性能是不一样的,这表明了模态之间天然的不同,也为前文提到的模态不平衡现象的存在提供了支撑,本文的方法比全部先进单模态识别方法都要好些,说明了其具备达成模态间协同融合的能力.

表 1 在 VAIS 数据集上 MM-NET 与单模态方法的比较

Tab. 1 Comparison of MM-NET with single-modality methods on the VAIS dataset

模型	模态	精确率	召回率	F1 分数	准确率
Vit ^[13]	Visible	0.811 5	0.829 1	0.814 2	0.847 0
	IR	0.581 3	0.579 4	0.574 7	0.653 9
SwinTransformer ^[16]	Visible	0.828 0	0.805 2	0.795 2	0.834 2
	IR	0.725 5	0.704 3	0.670 4	0.737 7
Resnet ^[14]	Visible	0.805 3	0.803 3	0.799 4	0.803 3
	IR	0.685 8	0.681 2	0.679 9	0.681 2
Mambabout ^[15]	Visible	0.830 1	0.848 1	0.848 1	0.870 7
	IR	0.741 7	0.686 4	0.697 3	0.748 6
MM-NET	Visible+IR	0.911 8	0.908 9	0.905 3	0.908 9

本文把自身采用的方法和当下最先进的多模态识别方法做了比较,这些方法有梯度调制方法、动态融合技术、模态特定测量策略,它们对当代多模态学习的主要研究方向意义重大.表 2 给出的结果显示,本文用的方法比对比方法有一定改进.能取得这一成果是因为本文方法能从多个不同角度评估模态贡献,这种能力超过了固定融合机制和局部模态测量技术.

表 2 在 VAIS 数据集上 MM-NET 与多模态方法的比较

Tab. 2 Comparison of MM-NET with multimodal in the VAIS dataset

模型	精确率	召回率	F1 分数	准确率	模型	精确率	召回率	F1 分数	准确率
QMF ^[19]	0.896 5	0.898 0	0.893 7	0.898 0	MMTM ^[21]	0.814 1	0.810 7	0.803 5	0.810 7
CAFN ^[18]	0.886 6	0.882 3	0.873 8	0.882 4	Transfuser ^[20]	0.863 3	0.858 5	0.855 0	0.858 5
OGM-GE ^[17]	0.879 8	0.874 0	0.854 4	0.874 0	MM-NET	0.911 8	0.908 9	0.905 3	0.908 9

3.4 消融实验

实验表明,MM-NET 优于最先进的多模态分类方法.性能提升归因于 MM-NET 的两个关键组件:密集 KAN 模块和贡献评分融合机制.为了进一步探索每个组件的贡献,本文进行了消融研究.

1)密集 KAN 模块的重要性:为了评估密集 KAN 模块的有效性,本文比较了有无该模块的 MM 的性能.在没有密集 KAN 模块的配置中,本文在网络中用标准 KAN 模块和传统 MLP 替换了它.如附录表 S1 所示,结果清楚地表明,配备密集 KAN 模块的 MM-NET 优于其对应配置.

2)贡献评分融合的重要性:为了进一步分析所提出的贡献评分融合机制的有效性,本文在保持网络主要骨干不变的情况下,仅在决策层替换了融合机制.本文将本文的方法与几种经典融合机制进行了比较:Sum-Fusion、Film、Concat 和 GateFusion.如附录表 S2 所示,本文的方法在所有性能指标上均优于这些方法.这归因于本文对不同模态的多方面测量,为每个模态提供贡献分数,并为后续决策级融合提供理论基础.

4 结 论

本文主要应对多模态船舶识别的过程中,不同模态的传感器数据所提供信息量的不同造成的模态不均衡问题,给出了基于模态度量的一种多模态融合识别方式,对于所捕捉的红外和可见光图像,制订了特征过滤模块,并通过结合新兴 KAN 所具备的优势,设计出一种有着密集连接的 Dense KAN 网络,来增强特征的非线性提取能力.在决策层引入模态度量评估方法,对多模态的图像特征予以融合,从而能够做到对目标的精确识别,通过结合 KAN 的观念,以及按照不确定性所做的模态评估,该模型为多模态识别给出了一套更准确、高效且容易理解的解决措施,本文提出的 MM-NET 于公开的 VAIS 数据集上,在准确性和计算效率上都优于现有的最先进方法,反映了本文所提出方法的优越性.

附录见电子版(DOI:10.16366/j.cnki.1000-2367.2025.06.25.0001).

参 考 文 献

[1] 郑海君,葛斌,夏晨星,等.多特征聚合的红外-可见光行人重识别[J].光电工程,2023,50(7):230136.
Zheng H J,Ge B,Xia C X,et al.Infrared-visible person re-identification based on multi feature aggregation[J].Opto-Electronic Engineering,2023,50(7):230136.

[2] Gao X Y,Shi Y B,Zhu Q,et al.Infrared and visible image fusion with deep neural network in enhanced flight vision system[J].Remote Sensing,2022,14(12):2789.

[3] Dong S K,Feng J F,Fang D X.A novel multiscale contrastive learning network for fine-grained ocean ship classification[J].IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2024,17:9989-10005.

[4] Xiao Q,Liu B,Li Z Y,et al.Progressive data augmentation method for remote sensing ship image classification based on imaging simulation system and neural style transfer[J].IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2021,14:9176-9186.

[5] He J L,Chang W L,Wang F P,et al.Multi-scale dense networks for ship classification using dual-polarization SAR images[C]//2023 IEEE Radar Conference (RadarConf23). San Antonio:IEEE,2023:1-6.

[6] 杨颖,杨艳秋,余本功.考虑视图可信度的用户多模态意图识别方法[J].电子与信息学报,2025,47(6):1966-1975.
Yang Y,Yang Y Q,Yu B G.Multimodal intent recognition method with view reliability[J].Journal of Electronics & Information Technology,2025,47(6):1966-1975.

[7] Zhang M M,Choi J,Daniilidis K,et al.VAIS:a dataset for recognizing maritime imagery in the visible and infrared spectrums[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).Boston:IEEE,2015:10-16.

[8] Liu J Y,Fan X,Huang Z B,et al.Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans:IEEE,2022:5792-5801.

[9] Tlig M,Bouchouicha M,Sayadi M,et al.Infrared-visible images' fusion techniques for forest fire monitoring[C]//2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). Sfax:IEEE,2022:1-6.

[10] Tang L F,Zhang H,Xu H,et al.Rethinking the necessity of image fusion in high-level vision tasks;a practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity[J].Information Fusion,2023,99:101870.

[11] Xie Z H,Zong S,Li Q,et al.Interactive residual coordinate attention and contrastive learning for infrared and visible image fusion in triple frequency bands[J].Scientific Reports,2024,14:90.

[12] Zhao Z X,Bai H W,Zhang J S,et al.CDDFuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Vancouver:IEEE,2023:5906-5916.

[13] Dosovitskiy A,Beyer L,Kolesnikov A,et al.An image is worth 16x16 words: transformers for image recognition at scale[EB/OL].[2025-06-10].https://arxiv.org/abs/2010.11929.

[14] He K M,Zhang X Y,Ren S Q,et al.Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas:IEEE,2016:770-778.

[15] Yu W H,Wang X C.MambaOut:do we really need mamba for vision [C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Nashville:IEEE,2025:4484-4496.

[16] Liu Z,Mao H Z,Wu C Y,et al.A ConvNet for the 2020s[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans:IEEE,2022:11966-11976.

[17] Peng X K,Wei Y K,Deng A D,et al.Balanced multimodal learning via on-the-fly gradient modulation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).New Orleans:IEEE,2022:8228-8237.

[18] He X Y,Wang Y,Zhao S,et al.Co-attention fusion network for multimodal skin cancer diagnosis[J].Pattern Recognition,2023,133: 108990.

[19] Zhang Q, Wu H, Zhang C, et al. Provable dynamic fusion for low-quality multimodal data[C]//International conference on machine learning. [S.l.]:PMLR,2023:41753-41769.

[20] Chitta K,Prakash A,Jaeger B,et al.TransFuser: imitation with transformer-based sensor fusion for autonomous driving[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2023,45(11):12878-12895.

[21] Vaezi Joze H R,Shaban A,Iuzzolino M L,et al.MMTM: multimodal transfer module for CNN fusion[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).[S.l.]:IEEE,2020:13286-13296.

Multimodal maritime ship recognition method based on modality metric

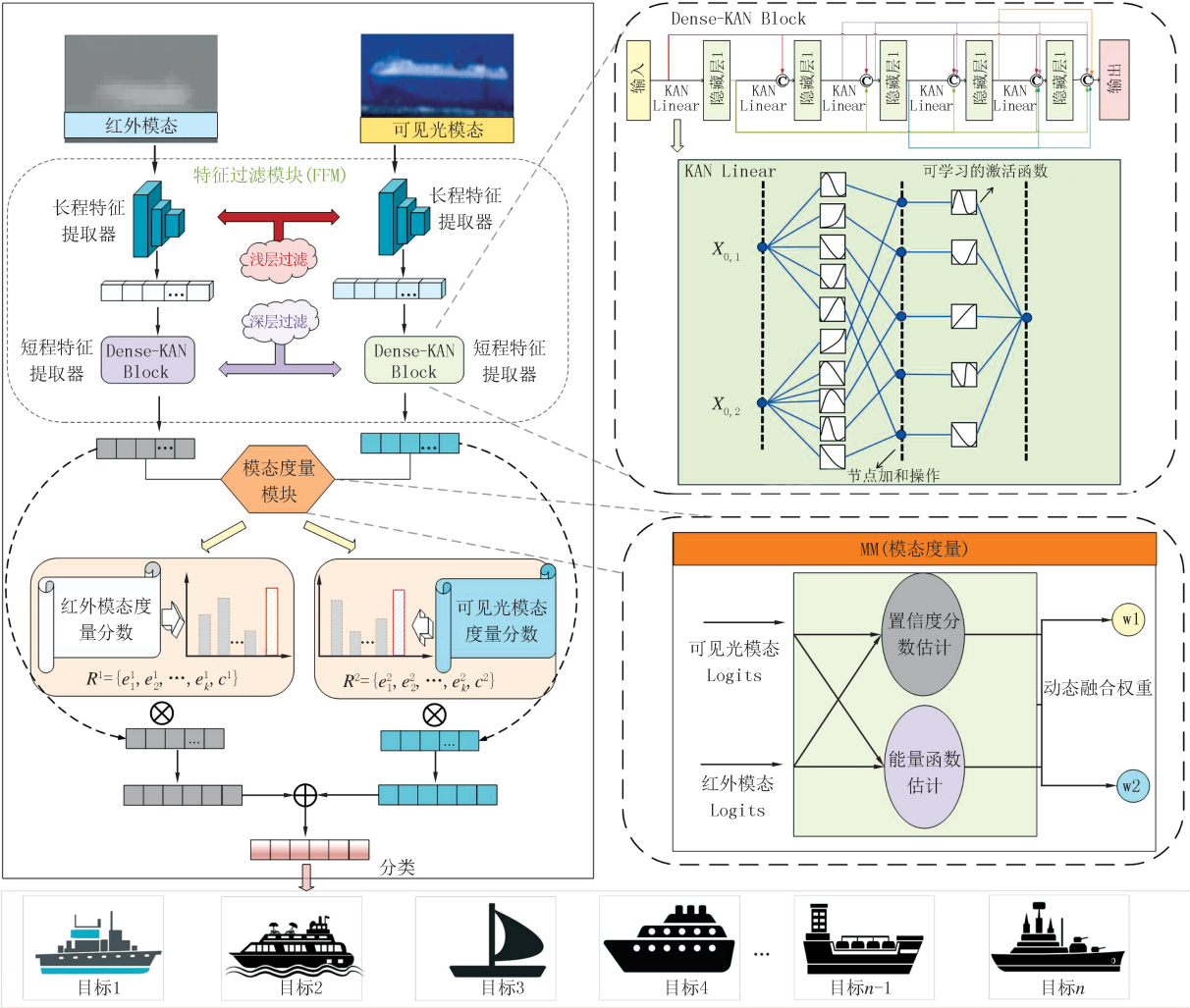
Xu Zhengwei^{1a,b}, Huang Peiji^{1a,b}, He Jianliang^{1a,b},Wang Shihao²

(1.a.College of Computer and Information Engineering(School of Artificifical Intelligence); b. Key Laboratory of Artificial Intelligence and Personalized Learning in Education of Henan Province, Henan Normal University, Xinxiang 453007, China; 2. College of Information and Communication Engineering, Harbin Engineering University, Harbin 150006, China)

Abstract: To address the shortcomings that current maritime multimodal recognition methods have in processing complex nonlinear features, effectively utilizing information from various modalities, and using fixed fusion strategies and weights, this paper puts forward a Modality Measurement-based Maritime Ship Recognition Network(MM-NET). The model has inside a Feature Filtering Module(FFM) which is used for effective suppression of useless information, reduction of distribution differences between different modalities, and improvement of the stability of multimodal feature picking work. In addition, this article devises a Modality Measurement(MM) mechanism which dynamically gives out fusion weights according to the contribution of every modal feature, therefore reducing to the smallest extent the influence that low-value modalities have on the final fusion performance. This mechanism can promote fusion effect degree on current scenarios and enviroment changes, and on self-adaption fusion weight changes. The method we put forward has carried on the benchmark test on the public VAIS maritime infrared-visible multimodal data set, thus it has shown that the accuracy and calculation efficiency outperform than the existing advanced methods.

Keywords: multimodal fusion; maritime ship recognition; feature filtering; modality metric

[责任编辑 陈留院 杨浦]



图S1 Diagram of the MM network

表 S1 Dense-KAN 模块有效性的消融实验

Tab. S1 Ablation study on the effectiveness of Dense-KAN Block

配置	精确率	召回率	F1 分数	准确率
Dense-KAN->KAN	0.895 6	0.892 5	0.880 9	0.892 5
Dense-KAN->MLP	0.891 6	0.894 4	0.898 6	0.894 4
本文的方法	0.911 8	0.908 9	0.905 3	0.908 9

表 S2 模态度量融合机制替换为其他融合机制的效果对比

Tab. S2 Comparison of the effects of replacing the modality metric fusion mechanism with other fusion mechanisms

配置	精确率	召回率	F1 分数	准确率	配置	精确率	召回率	F1 分数	准确率
MMFuse->Sum	0.824 3	0.817 9	0.812 4	0.817 9	MMFuse->Gate	0.871 0	0.861 6	0.857 6	0.861 6
MMFuse->Film	0.837 0	0.830 6	0.819 9	0.830 6	本文的方法	0.911 8	0.908 9	0.905 3	0.908 9
MMFuse->Concat	0.876 9	0.874 3	0.869 2	0.874 3					