HSE-Net：面向图像检索的谐波结构与边缘形态网络

熊杨 [1,*]
1. 重庆安全技术职业学院 人工智能学院，重庆 巴南 401320

摘要：针对复杂真实场景中背景杂乱、类间差异细微所导致的图像检索性能下降问题，现有深度检索模型多局限于单一空间域特征提取，不可避免地陷入“感受野困境”：扩大感受野会引入大量背景噪声，限制感受野又会丢失鲁棒的全局上下文。其根本原因在于，高频判别性边缘与随机背景噪声在空间域中深度耦合，难以有效解耦。为突破该瓶颈，本文提出一种谐波结构与边缘形态网络，基于“正交双域解耦”的新范式，将特征表示分解为两个物理正交的分支。在空间域，设计微边缘动态注意力模块（MEDA），利用方向梯度先验剥离平滑背景，精确定位目标的细粒度物理边缘；在频率域，引入宏谱上下文编码器（MSCE），将特征映射至傅里叶复谱空间，显式滤除高频杂波并提取高鲁棒性的全局宏拓扑结构；进一步提出正交域特征交互模块（ODFI），通过双向交叉注意力机制实现双域异构特征的深度协同与自适应融合。在 CUB-200-2011 和 Stanford Cars 两个细粒度公开数据集上的大量实验表明，HSE-Net 的检索性能显著优于现有主流方法，证明了所提方法在复杂检索任务中的优越性。

关键词：图像检索；正交双域解耦；特征解耦与融合；傅里叶变换；注意力机制

HSE-Net: Harmonic Structure and Edge Morphology Network for Image Retrieval

Yang Xiong[1,*]
1. College of Artificial Intelligence, Chongqing Vocational Institute of Safety Technology, Banan Chongqing 401320, China

Abstract: Due to severe background clutter and subtle inter-class differences, image retrieval in complex real-world scenarios remains challenging. Existing deep retrieval models are mostly limited to single spatial-domain feature extraction, inevitably falling into the receptive-field dilemma: enlarging the receptive field introduces substantial background noise, while restricting it sacrifices robust global context. Fundamentally, high-frequency discriminative edges and random background noise are deeply coupled in the spatial domain and cannot be easily decoupled. This paper proposes a Harmonic Structure and Edge Morphology Network (HSE-Net) under a novel orthogonal dual-domain decoupling paradigm. A Micro-Edge Dynamic Attention (MEDA) module leverages directional gradient priors to strip smooth backgrounds and precisely localize fine-grained physical edges. A Macro-Spectral Context Encoder (MSCE) maps features into the Fourier complex spectral space to filter high-frequency clutter and extract robust global macro-topology. An Orthogonal Domain Feature Interaction (ODFI) module fuses the two heterogeneous branches via bidirectional cross-attention. Extensive experiments on CUB-200-2011 and Stanford Cars show that HSE-Net outperforms state-of-the-art methods.

Keywords: Image retrieval; Orthogonal dual-domain decoupling; Feature decoupling and fusion; Fourier transform; Attention mechanism

1　引言

图像检索旨在从大规模图像数据库中高效检索出视觉与语义上相似的样本，在多媒体搜索、智能安防和目标识别等领域具有重要应用价值。然而，真实场景给图像检索带来了严峻挑战：背景杂乱、遮挡以及光照变化都会降低特征质量[1-3]；而全局外观高度相似的细粒度类别，则要求模型同时具备较强的全局鲁棒性与局部判别能力[4]。

现有深度模型大多在单一空间域中运行，存在固有的“感受野困境”——扩大感受野不可避免地引入背景噪声，缩小感受野又会丢失关键的全局上下文。从根本上看[5]，高频判别性边缘与高频背景噪声在空间域中深度耦合，二者极难分离[6-8]，这成为制约细粒度检索性能的核心瓶颈。

为突破这一瓶颈，本文从信号处理视角重新审视深度特征提取过程，提出一种谐波结构与边缘形态网络（HSE-Net），在“正交双域解耦”新范式下构建检索描述子。具体而言，设计微边缘动态注意力模块（MEDA），利用方向梯度先验隔离细粒度物理边缘；设计宏谱上下文编码器（MSCE），将特征映射至傅里叶域进行自适应噪声滤波与鲁棒宏拓扑提取；进一步提出具有双向交叉注意力的正交域特征交互模块（ODFI），在两个异构空间之间架起桥梁，实现深度协同融合。

2　HSE-Net 网络总体结构

为克服复杂图像检索场景中单一空间域的固有局限，本文提出谐波结构与边缘形态网络（HSE-Net）。如图 1 所示，HSE-Net 的核心思想是将特征提取过程解耦为“宏谱”与“微边缘”两个正交域，并通过跨域交互生成高鲁棒性的检索描述子。整个网络由四个主要部分组成：骨干特征提取模块、微边缘动态注意力（MEDA）模块、宏谱上下文编码器（MSCE）以及正交域特征交互（ODFI）模块。

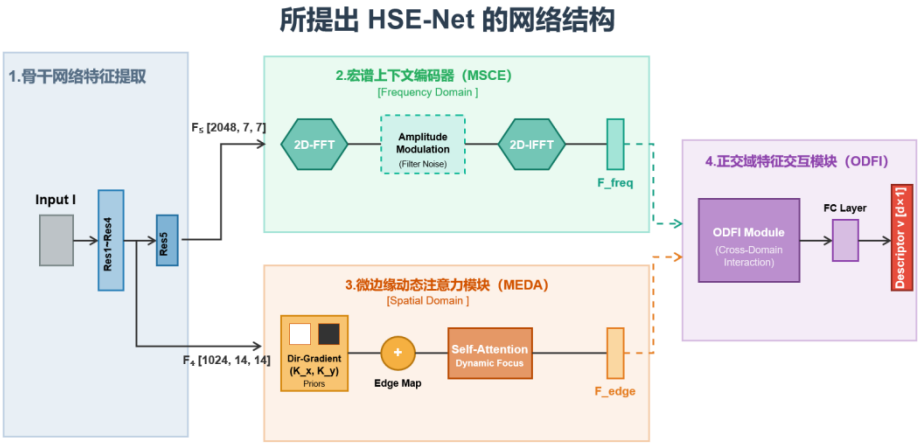


图 1　HSE-Net 网络总体架构

Fig.1　Overall Architecture of the Proposed HSE-Net

在特征提取的初始阶段，给定输入图像 I，本文以去除全连接层和全局平均池化层的 ResNet-50 作为基础骨干网络。输入图像首先经过 Res1 至 Res4 层，生成富含高频细节与局部语义的中层特征图 F4；随后 F4 继续经过 Res5 层，输出版本具有高层全局语义的深层特征图 F5。

为实现双域解耦，网络在该阶段分裂为对应空间域与频率域的双分支处理流程。空间域分支将中层特征 F4 送入 MEDA 模块，该模块利用水平和垂直梯度算子提取图像的物理边缘先验，并结合动态自注意力机制引导网络滤除背景杂波、强聚焦于目标的局部微观特征，输出空间边缘特征 F_edge。频率域分支将深层特征 F5 送入 MSCE 模块，通过二维快速傅里叶变换将空间特征映射至复谱空间，

在频域内执行自适应幅值调制以抑制高频噪声、增强低频宏结构，再经二维逆快速傅里叶变换映射回空间域，输出高鲁棒性的谱上下文特征 F_freq。

最后，空间细节特征 F_edge 与频率全局特征 F_freq 一同送入 ODFI 模块。该模块打破异构特征空间之间的壁垒，通过交叉注意力机制实现双域信息的优势互补与自适应对齐。融合后的特征经全连接层降维与 L2 归一化，最终生成 2048 维紧凑图像检索描述子 v，用于相似度度量。

3　关键模块设计

3.1　微边缘动态注意力模块（MEDA）

在传统图像检索网络中，局部分支通常采用空洞空间金字塔池化（ASPP）来提取多尺度特征。然而，ASPP 仅通过空洞率机械地扩大感受野，这种纯参数化映射缺乏显式的物理约束。当面对复杂背景杂波时，ASPP 容易将无关背景像素与目标细节混合，导致特征响应退化。为克服这一缺陷，本文设计微边缘动态注意力模块（MEDA），将“方向梯度先验”显式注入浅层网络，并结合动态注意力机制实现对微判别性边缘的高效聚焦。如图 2 所示，该模块的处理可分为两个核心阶段。

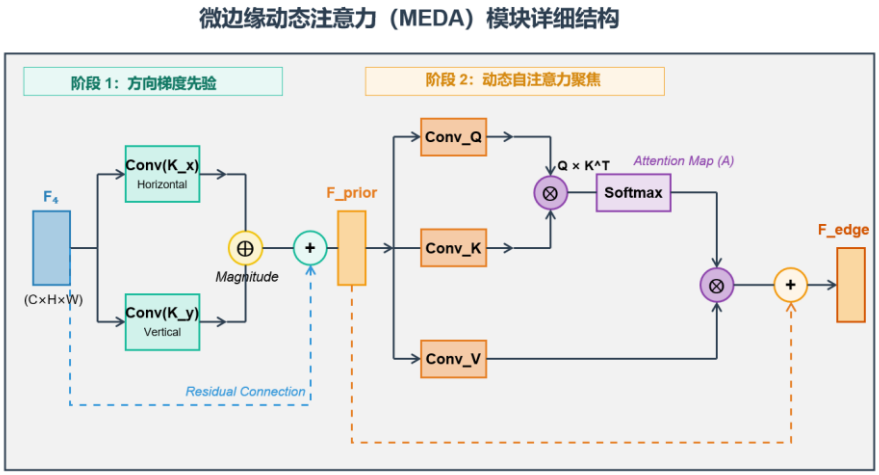


图 2　微边缘动态注意力（MEDA）模块详细结构

Fig.2　Detailed Architecture of the MEDA Module

在方向梯度先验提取阶段，自然图像中目标的结构信息高度集中于像素值剧烈变化的高频边缘区域。为使网络摆脱背景平滑区域的干扰，本文引入基于一阶差分梯度的空间算子。给定骨干网络输入的中间特征图 F4（C 为通道数，H、W 为空间维度），首先在特征空间中定义水平与垂直方向的方向梯度卷积核 K_x 和 K_y。为在保证梯度各向同性的同时降低计算复杂度，将经典 Sobel 算子扩展为深度卷积形式，权重参数固定且不参与反向传播。

利用 K_x 和 K_y 分别计算特征图在两个正交方向上的偏导数响应图 G_x 和 G_y，其中*表示深度卷积操作。随后计算局部梯度幅值图，并使用 1×1 逐点卷积进行通道重组，生成包含强物理边缘响应的先验特征。为防止梯度消失并保留原始语义特征，采用残差连接产生第一阶段输出特征 F_prior：其中|·|表示逐元素取绝对值，BN 表示批归一化，ReLU 为激活函数。经验证，F_prior 成功剥离了无关的低频背景，构建了目标的“高频物理骨架”。

在动态自注意力聚焦阶段，尽管 F_prior 强化了边缘，但在复杂图像（如树叶、波浪等背景）中仍存在大量高频边缘，会引起“假阳性”响应。为让网络自主判断“哪些边缘对检索真正有价值”，将 F_prior 送入动态自注意力机制。首先使用三个独立的 1×1 卷积层对 F_prior 进行线性投影，生成查

询矩阵 Q、键矩阵 K 和值矩阵 V，其中 N 为展平后的空间像素数，d 为降维后的通道数。接着在特征空间中计算稠密的亲和矩阵 A，其物理含义是度量图像中任意两个边缘像素之间的结构相关性：

其中 d_k 为缩放因子，用于防止点积运算引起的梯度爆炸。亲和矩阵 A 动态捕捉目标全局边缘之间的长程依赖。最后，将亲和矩阵 A 与值矩阵 V 进行矩阵乘法，完成空间特征的自适应重加权，并通过另一个残差连接输出最终的微边缘特征 F_edge。其中 λ 为可学习标量参数，初始化为 0，以保证网络在训练初期的稳定性。通过 MEDA 模块，网络彻底摆脱了传统 ASPP 对感受野的盲目依赖；在方向梯度先验与自注意力的双重约束下，F_edge 能够像显微镜一样极其精确地锁定目标的局部判别轮廓与纹理细节，为细粒度图像检索奠定坚实基础。

3.2 宏谱上下文编码器（MSCE）

在以往的图像检索方法中，为提取全局上下文信息，空间金字塔池化（SPP）或全局池化被广泛用于在空间维度上对深层特征进行下采样与聚合。然而，这种纯粹的“空间域聚合”存在一个致命的理论缺陷：感受野混叠。当面对复杂检索场景时，强烈的局部背景杂波（如突发亮斑、不规则遮挡）在空间域中表现为离散的高频噪声点，SPP 的滑动窗口与池化操作不可避免地将这些高频噪声与目标主体的宏观特征相混合，导致全局表示受到严重污染。

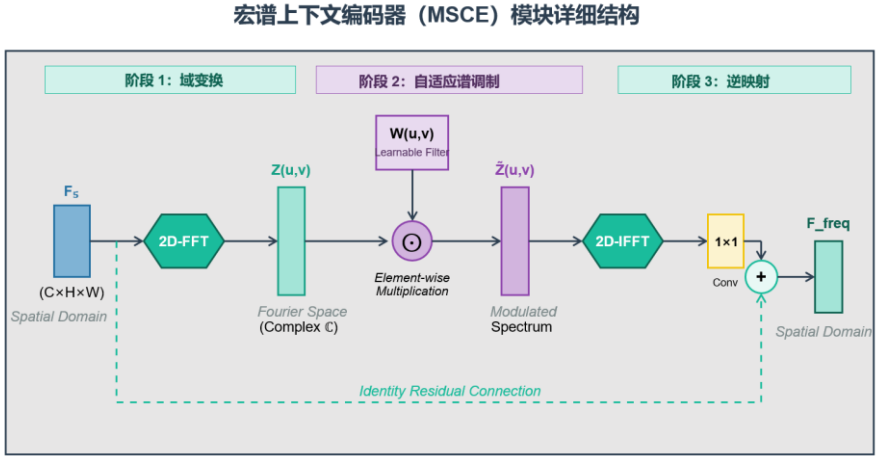


图 3 宏谱上下文编码器（MSCE）模块详细结构

Fig.3 Detailed Architecture of the MSCE Module

为从根本上解决这一问题，本文突破传统 CNN 的空间域局限，创新性地提出宏谱上下文编码器（MSCE）。如图 3 所示，该模块基于信号处理理论，将特征从空间域映射至傅里叶复谱空间，在频域中显式解耦并调制宏观结构与高频噪声，再逆映射回空间域。整个过程包含三个关键阶段。

在傅里叶域变换阶段，给定骨干网络输出的深层特征图 F5，将其视为 C 个独立的二维空间信号。为捕获全局感受野，沿空间维度 H、W 施加二维离散傅里叶变换。对第 c 个通道 F^c，其在频率坐标(u,v)处的复谱 Z(u,v)可由二维快速傅里叶变换（2D-FFT）计算得到。变换后整个特征映射为复张量 Z(u,v)。在复谱空间中，图像的空间特征被完全解耦：低频区域（谱中心）密集编码目标的宏观形状、平滑轮廓与全局上下文；高频区域（谱边缘）则主要由来自空间域的背景杂波、纹理与随机噪声构成。值得注意的是，频域表示天然具有“全局感受野”，因为谱中任意一点都包含整个空间域的信息。

在自适应谱调制阶段，为增强检索所需的宏观结构共性并衰减无用的高频噪声，本文在频域引入可学习的复权重张量（即谱滤波器）W(u,v)。与 SPP 中刚性的池化操作不同，MSCE 通过网络反

向传播以数据驱动的方式学习该滤波器的权重。将输入谱 Z(u,v)与可学习滤波器 W(u,v)进行逐元素 Hadamard 积，实现谱调制，得到调制后的纯净谱 $\tilde{Z}(u,v)$。从物理直觉上看，W(u,v)充当了一个通道级自适应带通/低通滤波器：在网络优化过程中，它会自发地在低频区域学习较大的权重幅值，以强化图像的宏观几何拓扑；同时在对应特定环境噪声的高频区域学习趋近于零的权重，从而在特征层面实现彻底的“去噪净化”。

在空间逆映射与残差聚合阶段，完成谱调制后，通过二维逆快速傅里叶变换（2D-IFFT）将净化后的复特征 $\tilde{Z}$ 重新映射回真实物理空间。考虑到逆变换后的 $\tilde{F}$ 是对全局结构的强抽象，为稳定深度网络中的梯度传播并确保原始语义信息不丢失，引入 1×1 卷积层进行通道融合，并采用残差连接机制生成最终的宏谱特征 F_freq，其中 Re(·)表示取复数实部。由此，MSCE 模块在不损失任何空间分辨率的前提下获得了极为纯净、鲁棒的全局上下文，完全替代了传统全局分支中存在信息损失风险的池化模块。

3.3 正交域特征交互模块（ODFI）

为融合异构的微边缘特征 F_edge 与宏谱特征 F_freq，本文提出具有双向交叉注意力与自适应门控的 ODFI 模块。将二者分别投影至公共子空间后，施加两条对称的注意力路径：“宏引导微”路径由宏谱特征生成注意力权重，对微边缘特征进行结构级筛选，使网络聚焦于与全局拓扑一致的可靠边缘；“微精炼宏”路径则由微边缘特征反向生成注意力权重，对宏谱特征进行细节校正，抑制模糊的轮廓并增强具有判别意义的局部区域。随后，门控网络生成通道级权重，对两路信息进行自适应加权，产生最终融合特征 F_fused。经过全局平均池化与全连接层后，即可得到紧凑的检索描述子 v。

4 实验与结果分析

为验证 HSE-Net 在复杂图像检索任务中的有效性，本文在两个具有挑战性的细粒度公开数据集上进行了大量实验。

4.1 数据集与实验设置

为全面评估模型捕获微观局部边缘与宏观全局特征的能力，本文舍弃简单的低分辨率数据集，采用以下两个被广泛使用的细粒度图像检索数据集：

CUB-200-2011：该数据集包含 200 种不同鸟类，共 11788 张图像。由于不同鸟类物种外观高度相似，且图像背景复杂（如树枝、天空），该数据集对模型的微观判别能力提出了极高要求。本文使用前 100 类（5864 张图像）进行训练，后 100 类（5924 张图像）进行测试。

Stanford Cars：该数据集包含 196 类不同品牌和年份的汽车，共 16185 张图像。汽车图像在视角和光照方面变化显著，适合测试模型在频率域中的宏观抗干扰能力。本文采用前 98 类（8054 张图像）训练，后 98 类（8131 张图像）测试。

本文采用图像检索领域最权威的标准评价指标 Recall@K（R@K）。具体而言，对于测试集中的每张查询图像，模型检索特征距离最近的 K 张图像，若其中至少有一张与查询图像属于同一类，则记为 1，否则记为 0。本文主要报告 R@1、R@5 和 R@10 的结果；此外，为全面评价整体排序质量，同时报告平均精度均值（mAP）。

4.2 对比实验

为全面评价 HSE-Net 的检索性能，本文在 CUB-200-2011 和 Stanford Cars 两个细粒度基准数据集上，将所提方法与若干主流图像检索及深度度量学习方法进行对比，包括基础 ResNet-50+三元组损失（Baseline）、多相似度损失（Multi-Similarity）[9]、Proxy-Anchor[10]以及近期提出的跨尺度检索

网络 MoDM[11]。表 1 给出了不同方法在 R@1、R@5、R@10 和 mAP 上的定量对比结果。

表 1 HSE-Net 与现有主流方法在 CUB-200-2011 和 Stanford Cars 数据集上的性能对比

Table 1 Performance Comparison Between HSE-Net and State-of-the-Art Methods on CUB-200-2011 and Stanford Cars Datasets

数据集	方法	R@1/%	R@5/%	R@10/%	mAP/%
CUB-200	Baseline(ResNet50)	61.4	75.2	82.8	58.7
CUB-200	Multi-Similarity	65.7	78.1	85.3	63.4
CUB-200	Proxy-Anchor	68.4	80.6	87.1	66.2
CUB-200	MoDM	70.2	81.9	88.5	68.1
CUB-200	HSE-Net(本文)	73.8	84.5	91.2	71.6
Cars	Baseline(ResNet50)	71.5	83.6	88.4	67.3
Cars	Multi-Similarity	76.2	87.1	91.5	71.8
Cars	Proxy-Anchor	79.5	89.4	93.2	75.4
Cars	MoDM	81.1	90.7	94.1	77.9
Cars	HSE-Net(本文)	84.6	93.2	96.5	82.3

如表 1 所示，本文提出的 HSE-Net 在两个具有挑战性的数据集上均取得了最优的检索性能。在 CUB-200-2011 鸟类数据集上，由于背景极为杂乱（如树枝、树叶干扰），传统 Baseline 方法存在严重的感受野混叠问题，R@1 仅为 61.4%。相比之下，HSE-Net 凭借 MSCE 模块在频率域强力滤除高频背景噪声、MEDA 模块动态聚焦微边缘的优势，将 R@1 提升至 73.8%，较 Baseline 大幅提升 12.4 个百分点。同样，在 Stanford Cars 数据集上，车辆外观高度相似、判别依赖于车灯和格栅等细微局部差异，HSE-Net 取得了 82.3%的 mAP，优于 Proxy-Anchor 和 MoDM 等强基线。这充分证明了“正交双域解耦”特征交互机制在处理细粒度、高相似度检索任务中的显著优势，不仅有效克服了背景杂乱，还大幅增强了相似类别之间的特征判别能力。

4.3 消融实验

为深入考察 HSE-Net 中各核心模块（MEDA、MSCE、ODFI）对整体检索性能的具体贡献，本文在 CUB-200-2011 和 Stanford Cars 数据集上进行了一系列渐进式消融实验。所有消融变体均采用与完整模型完全相同的优化策略和损失函数进行训练。表 2 详细列出了不同模块组合下的性能变化。

表 2 HSE-Net 各组件消融实验结果

Table 2 Ablation Results of Each Component in HSE-Net

模型	MEDA	MSCE	ODFI	CUB-200(R@1/mAP)/%	Cars(R@1/mAP)/%
模型 A(Baseline)	—	—	—	61.4/58.7	71.5/67.3
模型 B	√	—	—	66.8/64.5	76.1/72.0
模型 C	—	√	—	65.2/63.8	77.9/73.5
模型 D(无 ODFI)	√	√	—	70.1/68.2	81.4/77.8
HSE-Net(本文)	√	√	√	73.8/71.6	84.6/82.3

对比模型 A 与模型 B 可知，单独引入 MEDA 模块使 CUB-200 和 Stanford Cars 上的 R@1 分别提升 5.4 和 4.6 个百分点，验证了方向梯度先验与自注意力机制能够有效剥离平滑区域、强锁定对细粒度检索至关重要的局部物理边缘（如鸟喙纹理、车灯形状），从而显著增强特征的微观判别力。模型 A 与模型 C 的对比表明，引入 MSCE 模块在背景高度杂乱的 CUB-200 数据集上带来稳定的性能

增益；尤其在视角变化较大的 Stanford Cars 数据集上，R@1 跃升至 77.9%，说明将特征映射至傅里叶复空间并进行自适应滤波，能够有效丢弃空间域高频背景噪声，同时提取对空间形变高度鲁棒的全局宏上下文。更重要的是，尽管通过朴素拼接融合空间与频率特征的模型 D 已经优于单分支模型，但仍存在明显瓶颈；引入 ODFI 模块后（即完整 HSE-Net），性能较模型 D 再次取得显著突破，CUB-200 和 Cars 上的 mAP 分别提升 3.4 和 4.5 个百分点。这一现象有力地说明，空间边缘特征与频率域宏特征具有高度异构性，简单拼接会导致语义错位；而 ODFI 通过双向交叉注意力机制成功实现了“宏引导微、微精炼宏”的深度协同融合，最大程度释放了双域解耦框架的潜力。

5 结论

本文针对背景杂乱、类间差异细微的复杂场景，提出了一种新型的谐波结构与边缘形态网络（HSE-Net）。为解决传统单一空间域方法的感受野困境，本文提出正交双域解耦范式：微边缘动态注意力（MEDA）模块通过方向梯度先验在空间域精确捕获细粒度目标轮廓，宏谱上下文编码器（MSCE）在傅里叶谱域提取抗噪鲁棒的全局拓扑特征，正交域特征交互（ODFI）模块则通过双向交叉注意力有效融合异构双域特征、弥合语义鸿沟。在 CUB-200-2011 和 Stanford Cars 数据集上的大量实验验证了 HSE-Net 优于现有主流方法，证明了其在判别性与鲁棒性图像检索中的有效性。未来工作将探索将该双域解耦范式扩展至跨模态检索与开放集识别任务，并进一步压缩模型以适配边缘部署场景。

参考文献

[1] 王智贤,雷浩鹏,王明文,等.一种面向手绘草图的多阶段跨域图像检索方法[J].计算机应用研究,2025,42(12):3832-3840.

[2] 何旭阳.基于机械零件真实图像的三维模型检索方法研究[D].华中科技大学,2025

[3] 韩雪昆,苗夺谦,张红云,等.双分支多粒度局部对齐的实例级草图图像检索[J].模式识别与人工智能,2023,36(8):701-711.

[4] 卢伟龙,赵维伟,闫丽飞,等.基于多模态注意力机制的图像文本跨模态检索技术[J].微型电脑应用,2026,42(8):11-15+20.

[5] 郭争龙,陈若男,周昕,等.云环境中跨模态特征协同的加密图像检索方法[J].科学技术与工程,2026,26(19):8286-8297.

[6] 杜兰,吴庆森,郭昱辰,等.面向 SAR 图像的视觉语言大模型知识迁移方法[J].雷达学报(中英文),2026,15(3):996-1012.

[7] 王立辉,李元,刘泽峰,等.基于频域协同与多尺度门控的电力巡检图像检索方法[J/OL].计算机工程,1-15. https://doi.org/10.19678/j.issn.1000-3428.0260250.

[8] 陈润兰.工业场景下细粒度图像检索研究[D].广西师范大学,2026.

[9] Wang X, Han X, Huang W, et al. Multi-similarity loss with general pair weighting for deep metric learning[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, 2019: 5017-5025.

[10] Kim S, Kim D, Cho M, et al. Proxy anchor loss for deep metric learning[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2020: 3235-3244.

[11] Xia Y, Sharma D, Yuan Y, et al. MoDM: efficient serving for image generation via mixture-of-diffusion models[C]//Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2026, 1: 163-182.

基金项目：重庆市教育委员会科学技术研究计划项目（KJQN202504708），重庆安全技术职业学院校级教学改革项目《生成式人工智能融入信息技术课程体系的探索与实践》(AQJG24-13)，万州区社会科学重点课题《智媒时代青少年数字安全素养的体系化培育路径研究》（WZKT2025173）。

[1,*]**第 1 作者&通讯作者简介：**熊杨（1999—），女，硕士，重庆安全技术职业学院人工智能学院教师，研究方向为计算机视觉、图像检索、深度学习。E-mail: 13068300092@163.com。