

联合卷积神经网络与视觉变换器的室内场景识别

魏俊杰^{1,2}, 王 满^{1,2}, 王 晨^{2,3}

(1. 湖北省服装信息化工程技术研究中心;

2. 武汉纺织大学 计算机与人工智能学院;

3. 纺织服装智能化湖北省工程研究中心, 湖北 武汉 430200)

摘 要: 室内场景识别在智能家居、机器人导航等领域应用潜力巨大,但其精度提升仍面临技术瓶颈。卷积神经网络(CNN)具备提取低级特征与构建局部关联的天然优势,却因难以建模长距离依赖关系,导致室内场景识别性能受制约;而视觉变换器(ViT)擅长捕捉长距离依赖,可有效弥补这一缺陷。为此,提出一种融合 CNN 与 ViT 的室内场景识别方法,旨在充分结合两者优势以提升识别精度。首先,设计局部增强视觉变换器模块,通过捕获丰富的局部特征来增强 Agent-CSWin Transformer 的局部建模能力;其次,构建语义增强卷积神经网络模块,以删除最后一个分类层的 ResNet50 为编码器、卷积层为解码器,实现对室内场景语义信息的有效挖掘;最后,将上述两个模块与原始 ResNet50 模型融合,生成综合的室内场景特征表示,充分发挥 CNN 与 ViT 的协同优势。在 MIT-67、SUN-116 和 Places365-120 三个基准数据集上的实验结果显示,所提方法的识别准确率分别达到 87.84%、81.22% 和 63.77%;消融实验不仅验证了其整体有效性,而且证实了各模块的必要性。

关键词: CNN; ViT; 语义信息; 室内场景识别

DOI: 10.11907/rjdk.251601

扫描二维码阅读全文:



中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(XXXX)0XX-0001-09

Joint CNN and Vision Transformer for Indoor Scene Recognition

WEI Junjie^{1,2}, WANG Man^{1,2}, WANG Chen^{2,3}

(1. Engineering Research Center of Hubei Province for Clothing Information;

2. School of Computer Science and Artificial Intelligence, Wuhan Textile University;

3. Hubei Provincial Engineering Research Center for Intelligent Textile and Fashion, Wuhan 430200, China)

Abstract: Indoor scene recognition has great potential in smart home, robot navigation and other fields, but its accuracy improvement still faces technical bottlenecks. Convolutional neural network (CNN) has the natural advantage of extracting low-level features and building local associations, but it is difficult to model long-distance dependencies, resulting in restricted performance of indoor scene recognition; The Vision Transformer (ViT) is good at capturing long-distance dependence, which can effectively remedy this defect. Therefore, this paper proposes an indoor scene recognition method combining CNN and ViT, aiming to fully combine the advantages of both to improve the recognition accuracy. Firstly, the local enhanced vision converter module is designed to enhance the local modeling ability of Agent CSWin Transformer by capturing rich local features; Secondly, the semantic enhanced convolutional neural network module is constructed, and the ResNet50, which deleted the last classification layer, is used as the encoder and the convolutional layer is used as the decoder to effectively mine the semantic information of indoor scenes; Finally, the above two modules are fused with the original ResNet50 model to generate a comprehensive representation of indoor scene features, giving full play to the collaborative advantages of CNN and ViT. The experimental results on MIT-67, SUN-116 and Places365-120 benchmark datasets show that the recognition accuracy of the proposed method reaches 87.84%, 81.22% and 63.77% respectively; The ablation experiment not only verified the overall effectiveness of the system, but also confirmed the necessity of each module.

Key Words: CNN; ViT; semantic information; indoor scene recognition

收稿日期: 2025-08-31

基金项目: 国家自然科学基金项目(62301373); 湖北省自然科学基金项目(2023AFB279)

作者简介: 魏俊杰(1999-), 男, CCF 学生会会员, 武汉纺织大学计算机与人工智能学院硕士研究生, 研究方向为图像处理; 王满(2000-), 男, 武汉纺织大学计算机与人工智能学院硕士研究生, 研究方向为图像处理; 王晨(1993-), 女, 博士, 武汉纺织大学计算机与人工智能学院副教授、硕士生导师, 研究方向为图像处理、机器视觉。本文通讯作者: 王晨。

0 引言

室内场景识别是计算机视觉领域的一个重要研究方向,其主要目标为自动对场景进行分类^[1]。该技术在智能家居、机器人导航、安全监控等领域具有广阔的应用前景。然而,室内场景识别面临诸多挑战:首先,场景内部存在光照、布局和视角的巨大差异;其次,室内往往包含多种物体且存在复杂的空间关系,增加了分类难度。早期识别方法依赖人工特征,如尺度不变特征变换(Scale Invariant Features Transform, SIFT)^[2]和方向梯度直方图(Histogram of Oriented Gradient, HOG)^[3],在鲁棒性和泛化性方面存在不足。随后,局部二值模式(Local Binary Pattern, LBP)^[4]、视觉词袋(Bag of Visual Words, BoVW)^[5]等中层特征被提出,但仍难以捕获全局语义。

近年来,深度学习推动了该领域的发展,其中卷积神经网络(Convolutional Neural Network, CNN)在局部特征提取方面具有优势,而视觉变换器(Vision Transformer, ViT)凭借自注意力机制在建模全局依赖方面表现突出^[6-7]。如何融合二者优势,兼顾全局与局部特征建模,已成为当前室内场景识别研究的重要方向。

1 相关工作

1.1 基于CNN的场景识别

CNN通过多层卷积与池化操作,在局部特征提取方面展现出强大能力。典型的CNN架构包括AlexNet^[8]、VGG-Net^[9]、GoogLeNet^[10]和ResNet^[11]。为进一步提升场景识别效果,Wang等^[12]提出广义注意力池化(Global Average Pooling, GAP)机制,使模型能自动关注关键信息区域。Wu等^[13]设计了MetaObject-CNN,通过元对象特征增强场景表征。Zhu等^[14]和Zhou等^[15]分别利用特征连接和Places数据库进行场景建模,均取得了良好效果。Chen等^[16]和Sun等^[17]进一步将语义分割特征与CNN特征融合,改善了对物体关系的建模能力。Feng等^[18]提出Conv2NeXt模型,引入一些注意力模块来优化特征图,使其表现更加稳健。然而,CNN的局限在于难以建立长程依赖关系,导致其在复杂场景下的表现受限。

1.2 基于ViT的场景识别

ViT的提出为场景识别提供了新思路。Dosovitskiy等^[19]首次将Transformer应用于图像分类,基于此提出的ViT模型将图像划分为固定大小的块(patch)并利用自注意力机制实现全局建模。Ren等^[20]提出一种结合池化层与低复杂度注意力哈希的视觉Transformer模型,用于高效图像检索。Touvron等^[21]提出一种数据高效的图像变换器(Data-efficient Image Transformers, DeiT)模型,该模型通过引入知识蒸馏技术来指导Transformer模型训练。Wang等^[22]提出金字塔视觉变换器(Pyramid Vision Transformer,

PVT),利用多尺度特征提取方法来提高高分辨率图像的识别效率。Dong等^[23]提出一种名为CSWin Transformer的高效视觉Transformer模型,通过引入交叉形窗口自注意力机制和局部增强位置编码,解决了全局自注意力计算成本高和局部自注意力感受野受限的问题。Han等^[24]提出一种名为Agent Attention的新型注意力机制,旨在解决传统Transformer模型中Softmax注意力机制计算复杂度高的问题,同时保持其强大的表达能力。以上方法虽在建模全局依赖方面优于CNN,但在计算效率方面存在不足。

1.3 融合CNN与ViT的场景识别

CNN擅长局部细节提取,ViT善于捕捉全局语义,因此融合两者优势成为新趋势。已有研究尝试通过特征级或分支级融合实现互补。例如,Iqbal等^[25]提出的BTS-ST网络将Swin-Transformer融入基于CNN的U-Net中,以提升全局建模能力。Yan等^[26]提出的Hybrid Conv-ViT网络采用双分支网络架构融合空间特征和光谱特征,从语义分割的角度实现分类任务。Hayat等^[27]提出的EffNetV2-ViT模型结合了EfficientNetV2和ViT的优势,其中EfficientNetV2用于快速处理图像并提取初始特征,而ViT则对这些特征进行分类。然而,以上方法仍存在对语义信息利用不足或依赖外部标注的问题。为此,本文提出联合CNN与ViT的JCVT(Joint CNN and Vision Transformer)方法用于室内场景识别,主要贡献点为:①提出一种简单有效的局部增强视觉变换器模块(Local-Enhanced Vision Transformer Module, LEVTM)模块来提升Agent-CSWin捕捉室内场景图像局部特征的能力,同时保持高效的长程交互;②提出一种新颖的语义增强卷积神经网络模块(Semantic-Enhanced Convolutional Neural Network Module, SECNNM),与依赖外部语义分割或人工标注的方法不同,其以编码器—解码器的卷积结构内部挖掘语义信息,能在端到端训练过程中自动增强语义表示,从而减少对外部标注的依赖,显著提升语义特征的可靠性和鲁棒性;③采用三支并行提取特征并融合的策略,以分别保留并强化CNN的局部结构、ViT的全局依赖与语义分支的高层语义信息,使不同信息流的互补更直观且易于消融分析。在多个基准数据集上的实验结果表明,所提方法在精度与泛化能力方面相较现有方法具有更好的表现。

2 JCVT模型构建

如图1所示,JCVT模型引入多个创新组件,通过捕捉全局和局部场景特征来提升室内场景识别性能。首先,该模型引入LEVTM来弥补Agent-CSWinTransformer的局限性,基于全局注意力机制捕捉局部细节特征^[24];其次,提出SECNNM来充分利用场景图像中的对象级信息;再次,引入ResNet50作为卷积分支的主干网络,充分发挥其在局部结构建模和低级视觉特征提取方面的优势,为模型提供稳

定且判别性强的基础特征表示;最后,设计一个三支模块,高效整合来自不同特征流的信息,从而综合利用全局

场景上下文、局部细节特征和对象级语义来生成场景的全面表示。

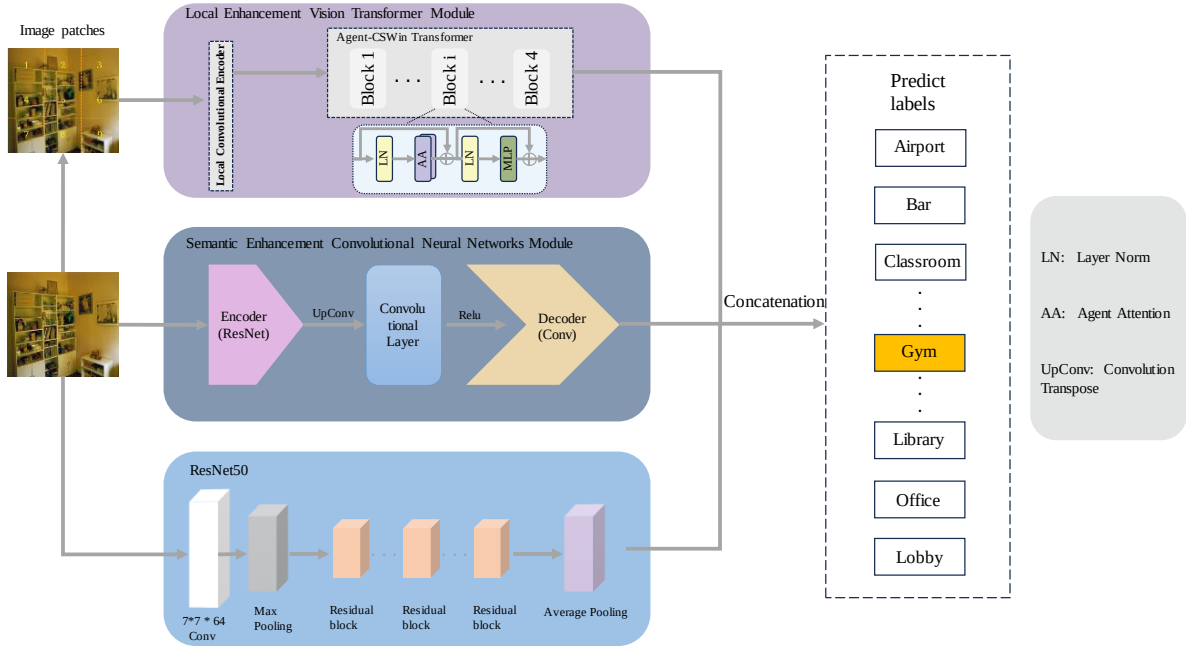


Fig. 1 JCVT model framework

图1 JCVT 模型框架

2.1 LEVTM

Agent-CSWin Transformer通过全局注意力机制有效捕获长距离依赖关系,但其在建模局部细节时存在不足,容易忽视物体的纹理、边缘等细粒度信息。在室内场景识别任务中,许多物品的判别依赖于局部结构(如桌角、门框、书架的层板等),单纯依赖全局信息往往会导致不同类别间的混淆。为此,本文在 Agent-CSWin Transformer 之前引入 LEVTM,以增强模型对输入图像的局部表征能力,其结构如图2所示。该模块在保持全局建模优势的同时,利用局部卷积编码器补充纹理和边缘等细节信息,从而实现局部与全局特征的有效互补。具体来说,LEVTM 首先通过

重叠卷积的令牌嵌入(Convolutional Token Embedding)将输入图像划分为图像块(Patch Token);其次在卷积令牌嵌入与 Agent-CSWin Transformer 之间设计一个由深度卷积与点卷积组成的局部卷积编码器,用于显式捕获局部依赖关系,最终获得更加丰富的局部特征表示。具体计算过程为:

$$\mathcal{X}_{i+1} = \text{Conv}_p(\text{Conv}_{p,G}(\text{DWConv}_{BN}(\mathcal{X}_i))) + \mathcal{X}_i \quad (1)$$

$$\mathcal{X}_o = \text{Conv}_p(\text{Conv}_{p,G}(\text{DWConv}_{3,BN}(\mathcal{X}_{i+1}))) + \mathcal{X}_{i+1} \quad (2)$$

式中: $\mathcal{X}_i$ 为补丁标记; $\text{Conv}_{p,G}$ 为逐点卷积; DWConv_{BN} 为深度卷积; $\text{DWConv}_{3,BN}$ 为3×3深度卷积; $\mathcal{X}_o$ 为输出局部特征。

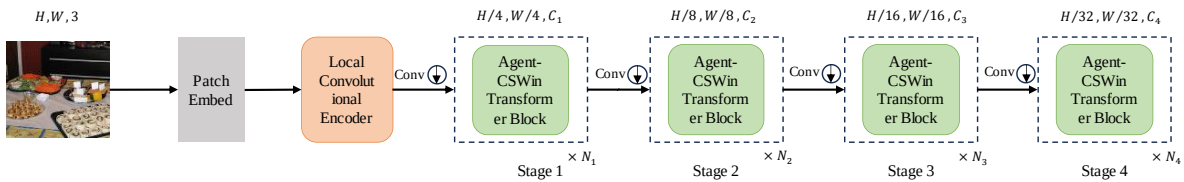


Fig. 2 Structure of LEVTM

图2 LEVTM 结构

为生成分层特征表示,输出的局部特征被输入到由4个层级(Stage)组成的 Agent-CSWin Transformer 网络中,每一层级对应不同尺度的特征表示,并通过逐级下采样实现空间分辨率的递减与通道维度的扩展,从而构建层次化的特征表达。为在相邻级之间进行传输,本文引入一个3×3卷积层,步长为2。该卷积层将空间维度(H和W)减半,同时将通道维度加倍。第i级的输出特征图包含(H/2i+1)×(W/2i+1)个token,与CSWin Transformer类似。由于全局注

意力机制计算成本较高,本文借鉴文献[24]中的方法,在传统的注意力三元组(Q, K, V)中引入一组额外的token q,形成四元组注意力范式(Q, q, K, V),简称为代理注意力(Agent Attention)。Agent Attention可表示为:

$$O^q = \text{Atten}^S(Q, q, \underbrace{\text{Atten}^S(q, K, V)}_{\text{Agent Aggregation}}) \quad (3)$$

Agent Broadcast

相当于:

$$O^q = \sigma(Qq^T)\sigma(qK^T)V \quad (4)$$

式中: q 为代理标记; $\sigma(\cdot)$ 为 Softmax 函数。

如式(3)所示,代理注意力机制包含两个 Softmax 注意力操作,分别为代理聚合(Agent Aggregation)和代理广播(Agent Broadcast)。首先,将代理标记 q 视为查询,在 q 、 K 与 V 之间计算注意力机制,从值中聚合出代理特征 Vq ;其次,将 q 作为键, Vq 作为值,与查询矩阵 Q 进行第二次注意力计算,从而将代理特征的全局信息广播到每个查询标记,最终输出 O^q 。代理注意力机制采用一种新的注意力范式,有效降低了原有注意力计算的复杂度。如此以来,LEVTM 既能增强输入特征的局部表示能力,又能保持长程交互能力,可有效补充 Agent-CSWinTransformer 模型的全局注意力机制,进而获得更丰富的场景特征表示。

2.2 SECNNM

传统 CNN 在局部模式识别方面表现优异,但其感受野有限,难以有效捕获全局上下文语义,尤其在复杂的室内场景中,不同物体之间的语义关系往往对场景判别具有关键作用,如“沙发+电视”更可能对应客厅,“书架+讲台”更可能对应教室。若缺乏这种语义层次建模,模型容易陷入对单一局部纹理的依赖,导致泛化性能不足。基于此,本文设计了 SECNNM。具体来说,本文首先移除最终的分层,并使用剩余的卷积层和池化层作为编码器;其次,使用反卷积层作为上采样模块,以提高特征图的分辨率;再次,上采样后的特征图被输入到标准卷积层,将通道数从 1 024 减少到 512;最后,选择 ReLU 激活函数,使用另一个卷积层作为解码器将特征图转换为语义类别。用公式表示为:

$$\mathcal{X}_{out} = \text{Decoder}\left(\text{Conv}_R\left(\text{UpConv}\left(\text{Encoder}\left(\mathcal{X}_{in}\right)\right)\right)\right) \quad (5)$$

式中: $\mathcal{X}_{in}$ 为输入特征; $\mathcal{X}_{out}$ 为输出语义类别; Conv_R 表示卷积操作后接 ReLU 激活函数; UpConv 为上采样层。

在该模块中,编码器保留了低级特征提取能力,而解码器则将特征图转化为高级语义特征,这种结构化设计有利于提取丰富的图像语义信息。在传统卷积分支中引入语义上下文建模机制可使卷积特征在捕获细节的同时具备更强的语义表达能力,从而弥补 CNN 在全局与语义关系建模方面的不足。

2.3 三支特征融合模块

室内场景识别任务往往同时依赖局部纹理、全局结构和跨对象语义关系,因此需要一种有效的多特征融合策略。受文献[28]的启发,本文设计了三支特征融合模块,在特征级别对来自 LEVTM、SECNNM 和 ResNet50 的特征进行拼接融合。首先,将原始图像分别输入到 3 个独立的分支进行特征提取;其次,对提取出的特征 $\mathcal{X}_1$ 、 $\mathcal{X}_2$ 和 $\mathcal{X}_3$ 进行整合;最后,将组合后的特征表示输入分类器进行分类。该过程可用公式表示为:

$$\mathcal{X}_1 = \text{LEVTM}(\mathcal{X}) \quad (6)$$

$$\mathcal{X}_2 = \text{ResNet50}(\mathcal{X}) \quad (7)$$

$$\mathcal{X}_3 = \text{SECNNM}(\mathcal{X}) \quad (8)$$

$$\mathcal{X}_4 = \text{Concat}(\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3) \quad (9)$$

式中: $\mathcal{X}_1$ 为 LEVTM 提取的特征; $\mathcal{X}_2$ 为 ResNet50 模型提取的特征; $\mathcal{X}_3$ 为 SECNNM 提取的特征; $\mathcal{X}_4$ 为最终融合特征; Concat 为连接函数。

通过联合优化,模型能同时保留局部细节、全局上下文与语义关系,从而实现更全面、更鲁棒的场景表示,最终提升分类准确率与泛化能力。

3 实验方法与结果分析

3.1 实验环境与参数设置

为保证实验结果的可复现性,本文对模型训练与测试所使用的软硬件环境及相关参数进行统一说明:①硬件环境。实验在单块 NVIDIA RTX 4090 GPU(32 GB 显存)上进行,能够满足模型训练的计算需求;②软件环境。模型基于深度学习框架 PyTorch 实现,运行于 Linux 操作系统, CUDA、cuDNN 版本与 GPU 环境相匹配;③输入设置。所有输入图像统一调整为 224×224 像素,以适配主流 CNN 与 ViT 的输入要求;④训练参数。批量大小(Batch Size)设置为 64,优化器采用 AdamW,权重衰减系数为 0.05,动量参数为 0.9;⑤训练策略。模型采用统一的训练策略,在不同数据集上保持一致的基本设置,以确保实验对比的公平性。

3.2 数据集

MIT-67 数据集共包含 15 620 张室内场景 RGB 图像,分为 67 个类别,每个类别 100~300 张图像,每个类别使用 80 张图像进行训练,20 张图像进行测试^[29]。由于类间相似度较高,且某些场景类别的图像中存在多个物体,分类任务挑战性较高。示例图像如图 3 所示。



Fig. 3 Examples of MIT-67 dataset

图 3 MIT-67 数据集示例图像

SUN-116 数据集基于大规模场景理解数据集 SUN-397 重新制作,本文从中选取 116 个室内场景类别,每个类别包含 100 幅图像,共计 11 600 幅图像^[30]。从每个场景类

别中随机选取 50 幅图像用于训练,另 50 幅用于测试。SUN-116 是一个标注精细的室内场景数据集,具体示例见图 4。



Fig. 4 Examples of SUN-116 dataset

图 4 SUN-116数据集示例图像

Places365-120 数据集基于最大的场景分类数据集 Places365 重新制作,本文从中选取 120 个室内场景类别,每个类别包含 200 幅图像,共计 24 000 幅图像^[15]。每个类别使用 160 幅图像进行训练,40 幅图像进行测试。该数据集的大多数图像包含大量复杂背景元素和遮挡情况,识别难度较大。具体示例见图 5。



Fig. 5 Examples of Places365-120 dataset

图 5 Places365-120数据集示例图像

3.3 实验方法

使用在 Places365 数据集上预训练的 ResNet50 模型来获取基于 CNN 的特征表示。使用在大规模 ImageNet 数据集上预训练的 Agent-CSWin Transformer 模型获得具有良好泛化能力的全局特征表示。SECNNM 用于提取语义特征表示。需要说明的是,ImageNet 与 Places365 数据集均作为预训练数据集使用,而本文实验主要在 MIT-67、SUN-116 和 Places365-120 三个室内场景数据集上进行。为减少随机性带来的影响,所有实验重复 10 次,最终结果取准确率的平均值。

3.4 实验结果与分析

3.4.1 与先进方法的比较

为验证所提方法的有效性,将其与 3 个数据集上的一

些最佳方法进行比较,包括基于 CNN 的方法(AlexNet^[8]、Dual CNN-DL^[31]、ResNet50^[11])、基于 ViT 的方法(Swin Transformer^[32]、TNT^[33]、Agent-CSWin^[24])、基于语义的方法(SRRM^[34]、Scene Essence^[35]、CCF^[36])和基于多分支的方法(DeepScene-Net^[37]、Multi-Scale CNN^[38]、SAS-Net^[39])。结果见表 1。为保证公平性,所有方法均采用相同的训练策略。具体而言,在 MIT-67 数据集上训练时,初始学习率为 0.000 1,dropout 率为 0.3;在 SUN-116 数据集上训练时,初始学习率为 0.000 5,dropout 率为 0.45;在 Places365-120 数据集上训练时,初始学习率为 0.000 1,dropout 率为 0.4。此外,所有方法均首先在 Place365 数据集上进行预训练;其次在 MIT-67 数据集上进行微调,以适应室内场景分类任务。

Table 1 Comparison of identification accuracy of each method

表 1 各方法识别准确率比较

方法	准确率/%		
	MIT-67	SUN-116	Places365-120
AlexNet	64.10	56.21	46.10
DeepScene-Net	71.00	64.42	49.20
Dual CNN-DL	76.56	69.96	52.53
Multi-Scale CNN	80.97	74.77	56.96
ResNet50	81.19	75.64	58.08
SWin-Transformer	81.42	73.38	56.42
SRRM	81.64	75.44	57.61
Scene Essence	83.92	76.72	60.01
TNT	84.33	79.28	60.15
Agent-CSWin	86.42	80.03	61.23
SAS-Net	87.10	80.80	63.32
CCF	87.30	81.03	63.54
JCVT	87.84	81.22	63.77

与 AlexNet、Dual CNN-DL 和 ResNet50 相比,JCVT 模型在 MIT-67 数据集上的识别准确率分别提升 23.74%、11.28% 和 6.65%;在 SUN-116 数据集上的准确率分别提升 25.01%、11.26% 和 5.58%;在 Places365-120 数据集上的准确率分别提升 17.67%、11.24% 和 5.69%。这是由于 JCVT 模型的混合架构克服了传统 CNN 在构建长距离依赖关系方面的局限性,同时保留了关键的局部视觉信息,实现了更全面的场景识别。

与 Swin Transformer、TNT 和 Agent-CSWin 相比,JCVT 模型在 MIT-67 数据集上的准确率分别提升 6.42%、3.51% 和 1.42%;在 SUN-116 数据集上的准确率分别提升 7.84%、1.94% 和 1.19%;在 PLACES365-120 数据集上的准确率分别提升 7.35%、3.62% 和 2.54%,展现出强大的跨数据集泛化能力。这是由于 JCVT 模型将 CNN 与 ViT 相结合,既保留了 ViT 在全局特征表示上的优势,又弥补了其在捕捉局部细节和计算效率上的不足。

与依赖外部语义标注的方法相比,JCVT 利用其内置的 SECNNM 自动捕获高级语义(例如对象共现关系),无需人工标注。在 MIT-67、SUN-116 和 Places365-120 数据集上,JCVT 的准确率相较 SRRM 分别提升 6.20%、5.78% 和

6.16%, 相较 Scene Essence 分别提升 3.92%、4.50% 和 3.76%。以上结果表明 JCVT 增强了提取高级语义表示的能力,并证实了深入探索场景特征与语义信息之间互补关系的可行性。CCF 方法融合了内容特征(基于图像中的背景与前景信息)和上下文特征(相似图像的标注),在 MIT-67、SUN-116 和 Places365-120 数据集上的识别准确率分别达到 87.30%、81.03% 和 63.54%。尽管该方法通过引入上下文信息提升了特征表达能力,但其性能仍略低于 JCVT 模型。这主要是由于该方法依赖外部文本标注信息,且对全局语义结构的建模能力有限,从而在刻画室内场景中复杂的对象关系和语义关联方面存在不足。

与 DeepScene-Net、Multi-Scale CNN 和 SAS-Net 相比, JCVT 在 MIT-67 数据集上的准确率分别提升 16.84%、6.87% 和 0.74%;在 SUN-116 数据集上的准确率分别提升 16.8%、6.45% 和 0.42%;在 PLACE365-120 数据集上的准确率分别提升 14.57%、6.81% 和 0.45%。值得注意的是,与依赖于多阶段训练的 SAS-Net 不同, JCVT 采用端到端框架,

联合优化语义和视觉特征,提高了模型的泛化能力。相较于其他多分支方法, JCVT 通过更加紧密的特征融合策略与端到端优化机制,实现了不同特征分支之间的有效协同,能够更充分地挖掘局部、全局与语义信息的互补性,从而在多个数据集上取得更优的识别性能。

总体来说, JCVT 在 3 个数据集上均取得最高准确率。值得一提的是,在极具挑战性的 Places365-120 数据集上, JCVT 的准确率比次优方法 CCF 高出 0.23%,证实了其对于多样化复杂场景类别的强大适应性。

为可视化每个类别的室内场景识别结果,图 6 展示了 JCVT 方法在 MIT-67 数据集上的混淆矩阵。可以看出,虽然 JCVT 在自助餐、赌场和计算机房等类别中达到最佳性能,但在礼堂、卧室和博物馆等类别中却具有较低的准确率。这是由于礼堂类别表现出明显的类内变异性;卧室类别中的一些图像被错误地归类到儿童房类别中;博物馆类别中的大多数图像被错误地归类到珠宝店类别中,这可能是由于类间相似性较高造成的。

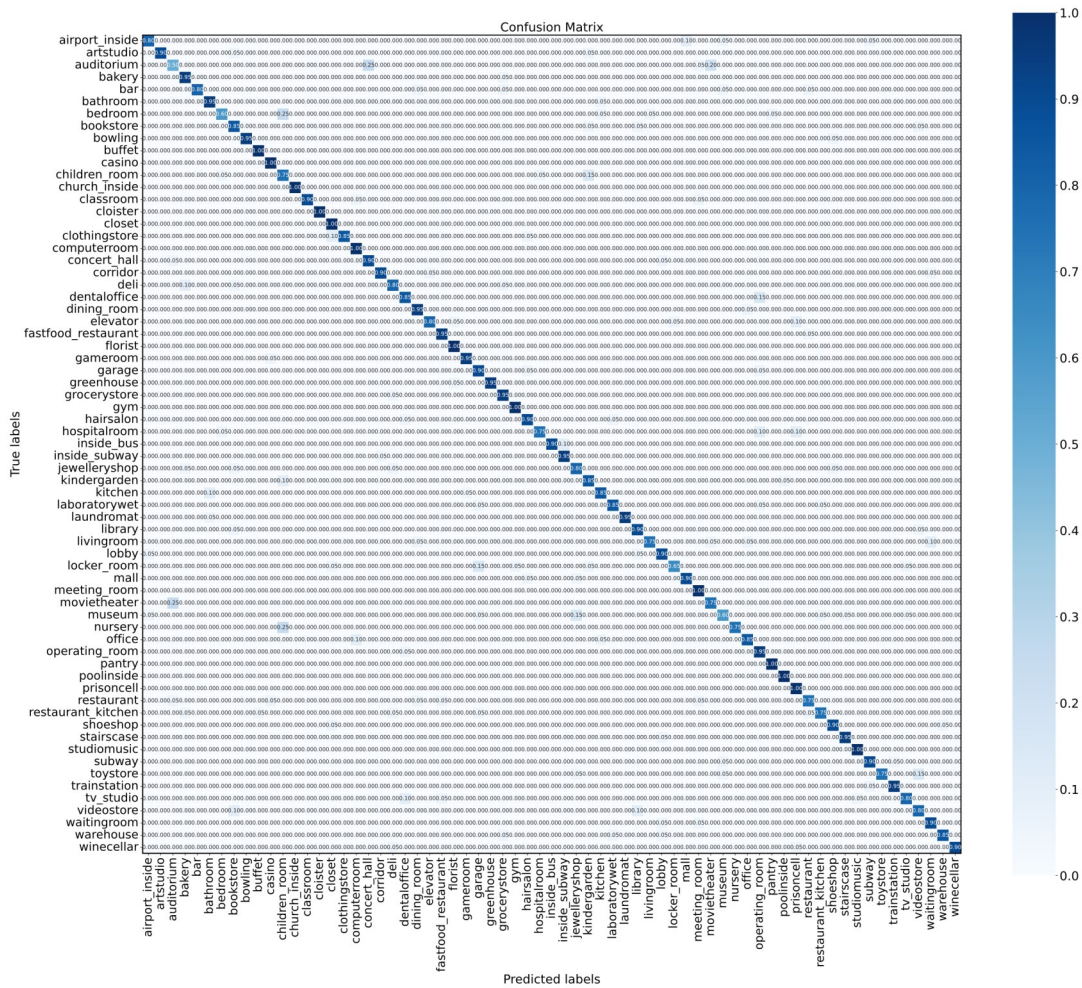


Fig. 6 Confusion matrix of the JCVT method on the MIT-67 dataset

图 6 JCVT 方法在 MIT-67 数据集上的混淆矩阵

3.4.2 消融实验

为评估所提模块的有效性,本文进行以下消融实验,每次仅移除一个模块,同时保持其他两个模块不变。实验

结果如表 2 所示。如第一行所示,移除 ResNet50 模块后,模型在 3 个数据集上的识别准确率相较完整模型分别下降 2.09%、1.03% 和 1.92%,验证了 ResNet50 模块的有效

性。如第二行所示,移除 SECNNM 后,模型在 3 个数据集上的识别准确率相较完整模型分别下降 0.75%、0.55% 和 0.83%,验证了 SECNNM 的有效性,也表明语义信息对于室内场景识别至关重要。如第三行所示,移除 LEVTM 后,模型在 3 个数据集上的识别准确率分别下降 0.83%、3.84% 和 2.58%,验证了 LEVTM 的有效性,表明建立图像中物体之间的长距离关系对于室内场景识别至关重要。此外,单独使用 CNN 模型(ResNet50)在 MIT-67 数据集上取得 85.91% 的准确率,优于 ViT 模型,说明 CNN 在局部纹理和结构特征提取方面具有优势。而单独使用 ViT 模型(Agent-CSWin)在 SUN-116 和 Places365-120 数据集上分别取得 76.94% 和 61.23% 的准确率,均优于 CNN,表明 ViT 在建模全局依赖和场景布局方面更具优势。然而,无论是单独使用 CNN 还是 ViT,其性能均明显低于完整模型 JCVT,这充分说明二者在特征表征上具有互补性。

Table 2 Ablation experiment results

表 2 消融实验结果

消融模型	准确率/%		
	MIT-67	SUN-116	Places365-120
LEVTM+SECNNM	85.75	80.19	61.85
LEVTM+ResNet50	87.09	80.67	62.94
SECNNM+ResNet50	87.01	77.38	61.19
ViT(Agent-CSWin)	84.42	76.94	61.23
CNN(ResNet50)	85.91	76.58	60.33
JCVT	87.84	81.22	63.77

3.4.3 参数分析

为探究不同参数对模型识别性能的影响,进行以下分析实验。JCVT 中只有一个特殊参数,即 LEVTM 中局部卷积编码器的卷积核大小。不同大小的卷积核会影响卷积层的感受野,较大的卷积核可以捕捉更大的空间特征,但

同时也会增加计算复杂度;而较小的卷积核则更注重细粒度的局部特征。因此,合适的卷积核大小可以提升室内场景识别效果。在 3 个数据集上,JCVT 模型在不同卷积核大小下的识别性能如表 3 所示。JCVT 在不同数据集上设置的卷积核大小存在差异,这主要是由于不同数据集在场景复杂度、目标尺度分布以及类间相似性等方面存在显著差异,对模型感受野的需求不同。可以看出,当卷积核大小分别设置为 31、15 和 25 时,JCVT 方法在 MIT-67、SUN-116 和 Places365-120 数据集上展现出最佳识别性能。

Table 3 Recognition performance of JCVT under different convolution kernel sizes

表 3 JCVT 在不同卷积核大小下的识别性能

数据集	卷积核大小	准确率/%
MIT-67	29	86.94
	31	87.84
	33	87.24
SUN-116	13	80.24
	15	81.22
	17	80.53
	23	63.35
Places365-120	25	63.77
	27	63.31

3.4.4 T-SNE 可视化

图 7 为 JCVT 在 3 个数据集上的 t-SNE 可视化结果,用于了解视觉特征分布。由于这 3 个数据集过于庞大,本文分别随机从 MIT-67、SUN-116 和 PLACE365-120 数据集中选取 10、20 和 20 个类别。可以看出,MIT-67 数据集中不同类别的图像相对容易区分;SUN-116 数据集中的图像更加紧凑,只有少数几个类别难以区分;PLACE365-120 数据集的判别性能较差,主要是由于其侧重于细节丰富的室内场景,类间相似性较高。

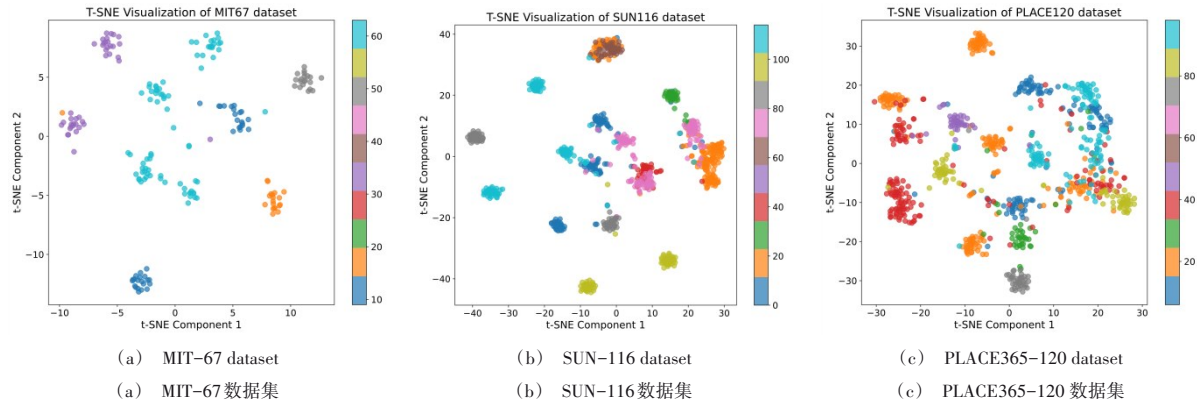


Fig. 7 t-SNE visualization results of JCVT on three datasets

图 7 JCVT 在 3 个数据集上的 t-SNE 可视化结果

此外,为将 JCVT 与基于 CNN 和基于 ViT 的方法进行视觉特征分布比较,本文选择 ResNet50 和 TNT 作为对照,在 MIT-67 数据集上进行实验,结果见图 8。可以看出,基于 CNN 的方法表现出明显的类内变异性,导致类内表示不够紧凑。基于 ViT 的方法虽然表现出更优的识别性能,但

对于某些类别的类间可分离性表现欠佳。相比之下,JCVT 在类内和类间均展现出较好的可分离性。

3.4.5 时间复杂度分析

为评估资源消耗水平与实时适用性,本文对不同方法在推理阶段的推理时间(Time)以及精度指标(AP、AR、

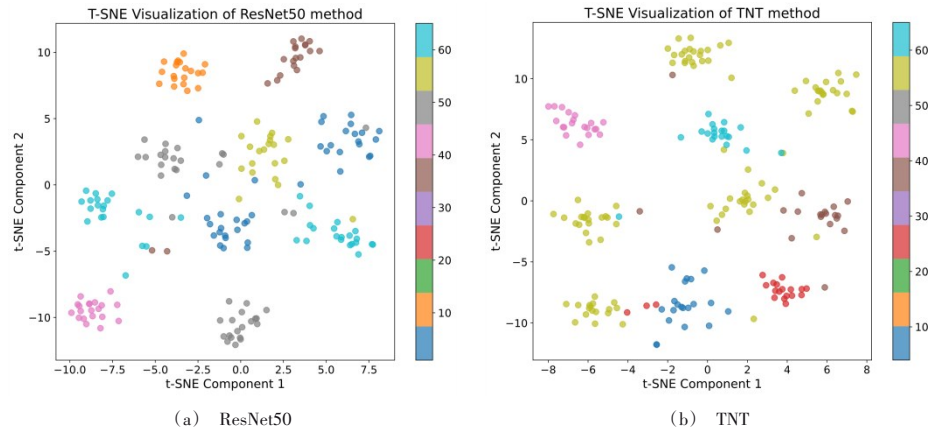


Fig. 8 t-SNE visualization results of ResNet50 and TNT

图8 ResNet50和TNT的t-SNE可视化结果

F1)进行比较。结果见表4、表5、表6。其中,Time指模型在推理阶段处理整个测试集所需的总时间,直接反映了算法的计算效率与实时响应能力;AP(Average Precision)即平均精度,用于衡量模型在各场景类别上的识别准确性;AR(Average Recall)即平均召回率,用于衡量模型正确识别出所有正样本的能力;F1为精度与召回率的调和平均数,是模型分类性能的综合评价指标。

Table 4 Performance of different methods on MIT-67 dataset

表4 不同方法在MIT-67数据集上的表现

方法	Time/s	AP/%	AR/%	F1/%
AlexNet	2.18	68.0	64.1	62.8
ResNet50	2.20	87.1	81.2	80.3
SWin-Transformer	2.21	85.6	81.4	79.8
Agent-CSWin	3.24	92.1	86.4	84.6
JCVT	2.73	93.4	87.8	87.6

Table 5 Performance of different methods on the SUN-116 dataset

表5 不同方法在SUN-116数据集上的表现

方法	Time/s	AP/%	AR/%	F1/%
AlexNet	11.23	57.8	56.2	55.2
ResNet50	13.23	79.7	75.6	73.3
SWin-Transformer	13.51	76.2	73.3	71.6
Agent-CSWin	15.25	83.5	80.0	78.8
JCVT	14.3	85.7	81.2	80.0

Table 6 Performance of different methods on the PLACES365-120 dataset

表6 不同方法在PLACES365-120数据集上的表现

方法	Time/s	AP/%	AR/%	F1/%
AlexNet	2.06	47.8	46.1	44.6
ResNet50	3.07	59.7	58.0	57.2
SWin-Transformer	3.06	57.6	56.4	55.7
Agent-CSWin	4.09	63.3	61.2	60.3
JCVT	3.33	65.2	63.7	62.8

可以看出,轻量级CNN(AlexNet)具有最低的推理时间,例如在MIT-67数据集上仅需2.18 s,在Places365-120数据集上为2.06 s,说明其计算复杂度最低,但其精度指标(AP、AR、F1)均明显低于其他方法,表明其表达能力不足。ResNet50相较AlexNet推理时间略有增加,但精度提

升显著。Transformer系列方法(Swin-Transformer, Agent-CSWin)的推理时间普遍高于CNN类方法,例如Agent-CSWin在SUN-116数据集上的推理时间达15.25 s,在Places365-120数据集上为4.09 s,计算复杂度显著增加,这主要是由于其自注意力机制需要进行全局特征建模。JCVT的推理时间介于以上两类方法之间,在MIT-67、SUN-116和Places365-120数据集上的推理时间分别为2.73 s、14.3 s和3.33 s。单独的CNN方法在3个数据集上的F1分数为44.6%~80.3%,总体不及单独的ViT方法,而JCVT的精度指标最优。总体来说,JCVT在保证识别精度的同时降低了计算复杂度,体现出较强的实用价值。

4 结语

针对室内场景识别方法难以兼顾全局上下文信息关联与局部细微特征的问题,本文提出一种结合CNN与ViT的室内场景识别方法。该方法通过构建多分支协同建模机制,实现了不同层次特征的有效互补与统一表达,提升了模型对复杂场景的表征能力。在多个室内场景数据集上的实验结果表明,所提方法在识别性能与泛化能力方面均优于现有方法,验证了其设计的有效性。然而,该方法仍存在一定局限性:一是其依赖监督学习,需要大量标注样本;二是目前实验主要在标准公开数据集上完成,与真实复杂环境仍存在差距。未来工作将从以下两个方向展开:一是探索少样本或零样本条件下的室内场景识别方法,以降低对大规模标注数据的依赖;二是针对光照变化、遮挡和动态物体干扰等真实场景因素开展更具挑战性的验证实验。具体而言,可以通过数据增强手段模拟光照与遮挡条件、通过跨数据集验证评估模型的泛化性以及在实际室内视频/图像序列上进行评测,以系统检验模型在实际应用环境下的鲁棒性和适应性,这将有助于推动其在智能家居、安防监控和人机交互等应用场景中的落地。

参考文献:

- [1] CHEN K, LAI Y K, HU S M. 3D indoor scene modeling from RGB-D data:

- a survey[J]. *Computational Visual Media*, 2015, 1(4): 267–278.
- [2] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. *International Journal of Computer Vision*, 2004, 60(2): 91–110.
 - [3] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005: 886–893.
 - [4] GUO Z, ZHANG L, ZHANG D. A completed modeling of local binary pattern operator for texture classification[J]. *IEEE Transactions on Image Processing*, 2010, 19(6): 1657–1663.
 - [5] YANG J, JIANG Y G, HAUPTMANN A G, et al. Evaluating bag-of-visual-words representations in scene classification[C]//Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval, 2007: 197–206.
 - [6] AFIF M, AYACHI R, SAID Y, et al. Deep learning based application for indoor scene recognition[J]. *Neural Processing Letters*, 2020, 51(3): 2827–2837.
 - [7] XIE L, LEE F, LIU L, et al. Scene recognition: a comprehensive survey[J]. *Pattern Recognition*, 2020, 102: 107205.
 - [8] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. *Communications of the ACM*, 2017, 60(6): 84–90.
 - [9] WANG L, GUO S, HUANG W, et al. Places205-vggnet models for scene recognition[DB/OL]. <https://arxiv.org/abs/1508.01667>.
 - [10] TANG P, WANG H, KWONG S. G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition[J]. *Neurocomputing*, 2017, 225: 188–197.
 - [11] WU Z, SHEN C, VAN DEN H A. Wider or deeper: revisiting the resnet model for visual recognition[J]. *Pattern Recognition*, 2019, 90: 119–133.
 - [12] WANG Y, ZHOU W, ZHANG Q, et al. Convolutional neural networks with generalized attentional pooling for action recognition[C]//2018 IEEE Visual Communications and Image Processing, 2018: 1–4.
 - [13] WU R, WANG B, WANG W, et al. Harvesting discriminative meta objects with deep CNN features for scene classification[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1287–1295.
 - [14] ZHU Y, NEWSAM S. Densenet for dense flow[C]//2017 IEEE International Conference on Image Processing, 2017: 790–794.
 - [15] ZHOU B, LAPEDRIZA A, KHOSLA A, et al. Places: a 10 million image database for scene recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 40(6): 1452–1464.
 - [16] CHEN B X, SAHDEV R, WU D, et al. Scene classification in indoor environments for robots using context based word embeddings[DB/OL]. <https://arxiv.org/abs/1908.06422>.
 - [17] SUN N, LI W, LIU J, et al. Fusing object semantics and deep appearance features for scene recognition[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018, 29(6): 1715–1728.
 - [18] FENG J, TAN H, LI W, et al. Conv2NeXt: reconsidering Conv NeXt network design for image recognition[C]//2022 International Conference on Computers and Artificial Intelligence Technologies, 2022: 53–60.
 - [19] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[DB/OL]. <https://arxiv.org/abs/2010.11929>.
 - [20] REN H, GUO J, CHENG S, et al. Pooling-based visual Transformer with low complexity attention hashing for image retrieval[J]. *Expert Systems with Applications*, 2024, 241: 122745.
 - [21] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//International Conference on Machine Learning, 2021: 10347–10357.
 - [22] WANG W, XIE E, LI X, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 568–578.
 - [23] DONG X, BAO J, CHEN D, et al. Cswin transformer: a general vision transformer backbone with cross-shaped windows[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 12124–12134.
 - [24] HAN D, YE T, HAN Y, et al. Agent attention: on the integration of softmax and linear attention[DB/OL]. <https://arxiv.org/abs/2312.08874>.
 - [25] IQBAL A, SHARIF M. BTS-ST: swin transformer network for segmentation and classification of multimodality breast cancer images[J]. *Knowledge-Based Systems*, 2023, 267: 110393.
 - [26] YAN H, ZHANG E, WANG J, et al. Hybrid Conv-ViT network for hyperspectral image classification[J]. *IEEE Geoscience and Remote Sensing Letters*, 2023, 20: 1–5.
 - [27] HAYAT M, AHMAD N, NASIR A, et al. Hybrid deep learning efficient-netv2 and vision transformer (effnetv2-vit) model for breast cancer histopathological image classification[J]. *IEEE Access*, 2024, 12: 184119–184131.
 - [28] WANG C, PENG G, DE BAETS B. Deep feature fusion through adaptive discriminative metric learning for scene recognition[J]. *Information Fusion*, 2020, 63: 1–12.
 - [29] QUATTONI A, TORRALBA A. Recognizing indoor scenes[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009: 413–420.
 - [30] XIAO J, HAYS J, EHINGER K A, et al. Sun database: large-scale scene recognition from abbey to zoo[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010: 3485–3492.
 - [31] LIU Y, CHEN Q, CHEN W, et al. Dictionary learning inspired deep network for scene recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 7178–7185.
 - [32] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012–10022.
 - [33] HAN K, XIAO A, WU E, et al. Transformer in transformer[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 15908–15919.
 - [34] SONG C, MA X. Srrm: semantic region relation model for indoor scene recognition[C]//2023 International Joint Conference on Neural Networks, 2023: 1–8.
 - [35] QIU J, YANG Y, WANG X, et al. Scene essence[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 8322–8333.
 - [36] SITAULA C, ARYAL S, XIANG Y, et al. Content and context features for scene image representation[J]. *Knowledge-Based Systems*, 2021, 232: 107470.
 - [37] YEE P S, LIM K M, LEE C P. DeepScene: scene classification via convolutional neural network with spatial pyramid pooling[J]. *Expert Systems with Applications*, 2022, 193: 116382.
 - [38] HERRANZ L, JIANG S, LI X. Scene recognition with CNNs: objects, scales and dataset bias[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 571–579.
 - [39] LÓPEZ-CIFUENTES A, ESCUDERO-VINOLO M, BESCÓS J, et al. Semantic-aware scene recognition[J]. *Pattern Recognition*, 2020, 102: 107256.

(责任编辑:尹晨茹)