AI-DETR:自适应加权的可解释目标检测方法

鲁银圆[1],许升全[2],谢娟英[1*]
(1. 陕西师范大学计算机科学学院,陕西西安 710119;2. 陕西师范大学生命科学学院,陕西西安 710119)

摘 要: 检测变换器(DEtection TRansformer,DETR)是计算机视觉和多模态学习等领域的研究热点,但其解码器学习偏差存在层间传递,且不同层交叉注意力计算使用相同参考点、编码器输出特征语义模糊,严重影响模型性能. 本文针对DETR的上述缺陷,以Conditional DETR为基线模型,将交叉注意力机制解耦为权重和值向量两部分,提出层间自适应注意力权重更新(Inter-layer Adaptive Attention Weight Refinement,IAAWR)方法,动态调节解码器不同层的交叉注意力权重,削弱学习偏差层间传递;提出值向量自适应增强(Adaptive Feature Enhancement,AFE)方法,采用分治思想改善编码器各层对目标局部区域的特征提取能力,显著增强输出特征的语义性;提出无参数迭代矫正预测框参考点(Iterative Reference Point Refinement,IRPR)方法,实现预测框参考点动态更新,增强回归预测的灵活性和精细度. 融合以上三个创新点改进基线模型Conditional DETR,得到自适应的可解释目标检测变换器(Adaptive and Interpretable DETR,AI-DETR). 新模型AI-DETR仅增加了11个可学习参数,其平均精度(Average Precision,AP)指标在公开数据集MS-COCO(MicroSoft Common Objects in COntext)上比基线模型Conditional DETR提升1.8个百分点,在更具挑战性的野外环境下蝴蝶数据集Butterfly_2018和Butterfly_2023上分别提升1.3个百分点和0.8个百分点. 通过定性、定量分析及结果可视化,详细阐述和论证了AI-DETR模型各创新点的具体贡献.

关键词: 目标检测;蝴蝶检测;深度学习;检测变换器(DETR)模型;交叉注意力机制
基金项目: 国家自然科学基金(No.62076159)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)07-2279-26
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250038

AI-DETR: Interpretable Object Detection Method Based on Adaptive Weighting

LU Yin-yuan[1], XU Sheng-quan[2], XIE Juan-ying[1*]
(1. *School of Computer Science*, *Shaanxi Normal University*, *Xi'an*, *Shaanxi* 710119, *China*;
2. *Collegel of Life Sciences*, *Shaanxi Normal University*, *Xi'an*, *Shaanxi* 710119, *China*)

Abstract: Detection transformer (DETR) has been emerging as a hotspot in computer vision, multimodal learning and other fields. However, its performance is heavily affted by the learning feature bias transmission between decoder layers, and the same reference points used by the cross-attention of different decoder layers, and the semantic vagueness of the encoder output features. To address these deficiencies, this paper employs Conditional DETR as the baseline and decouples its cross-attention mechanism into weights and values, then proposes an inter-layer adaptive attention weight refinement (IAAWR), with the aim of dynamically adjusting the cross-attention weights of different layers of the decoder, with a review to weakening the inter-layer transfer of learning bias. In addition, an adaptive feature enhancement (AFE) method is proposed utilizing divide and conquer idea, with the aim of improving the feature extraction capability of each layer of the encoder for the local region of the target, resulting in the enhancement of semantics in the output features. Furthermore, the strategy of parameter-free iterative reference point refinement (IRPR) is proposed to achieve dynamic update of the reference points of the prediction box, enhancing the flexibility and fineness of regression prediction. These three innovations have been integrated into the baseline model Conditional DETR, resulting in an adaptive and interpretable DETR model referred to adaptive and interpretable DETR (AI-DETR). This AI-DETR defeats the Conditional DETR in terms of average precision (AP) on the publicly available dataset microsoft common objects in context (MS-COCO) with 1.8 percentage

收稿日期:2025-01-09;录用日期:2025-04-30;责任编辑:覃怀银
*通讯作者:谢娟英

points and on the very challenging real-world datasets Butterfly_2018 and Butterfly_2023 datasets with 1.3 and 0.8 percent points, respectively. The qualitative and quantitative analyses, in conjunction with visualisations of the results, elucidate and validate the individual contribution of each innovation within the AI-DETR.

Key words: object detection; butterfly detection; deep learning; detection transformer (DETR); cross-attention mechanism
Foundation Item(s): National Natural Science Foundation of China (No.62076159)

1 引言

目标检测是计算机视觉领域的一项基础任务，是影响目标跟踪、行人重识别、视频理解等多类计算机视觉下游任务性能的核心. 基于卷积神经网络的目标检测方法近年取得了长足发展[1~4]. 然而此类方法需要人工组件，如锚框、非极大值抑制等，才能达到良好性能. 检测变换器(DEtection TRansformer，DETR)[5]基于集合匹配思想，采用卷积神经网络和Transformer[6]模型串联提取特征，在无锚框和非极大值抑制情况下，实现了与快速区域卷积神经网络(Faster Region-based Convolutional Neural Networks，Faster RCNN)[2]相当的检测效果. DETR需要训练更多轮次才可有效收敛，主要是由于Transformer存在参数利用率低和解码器中查询集学习难度大等问题. 为此，有不少针对DETR收敛速度慢的研究工作[7~11]. 其中，Conditional DETR引入空间条件查询集，将空间和语义特征解耦，优化Transformer交叉注意力计算过程，提升模型收敛速度，Conditional DETR的解决方案，被不少基于DETR的改进工作采用.

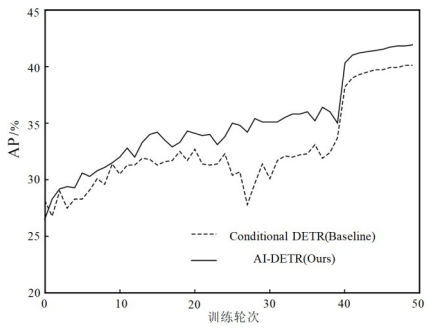
本文将Conditional DETR作为基线模型，解耦其交叉注意力计算过程，针对其解码器学习偏差存在的层层传递影响注意力权重，提出层间自适应注意力权重更新(Inter-layer Adaptive Attention Weight Refinement，IAAWR)方法，避免偏差传递，提升交叉注意力计算稳定性；针对其编码器最后输出的值向量存在信息丢失，提出值向量自适应增强(Adaptive Feature Enhancement，AFE)方法，动态融合编码器不同层输出，提升编码器值向量语义信息；针对解码器各层注意力使用相同参考点，影响预测框精准度，提出无参数迭代矫正预测框参考点(Iterative Reference Point Refinement，IRPR)方法，辅助模型动态改良参考点. 将该三个创新点融入基线模型Conditional DETR得到自适应的可解释目标检测变换器(Adaptive and Interpretable DETR，AI-DETR)新模型. 该AI-DETR模型仅引入11个可学习参数，在公开数据集MS-COCO(MicroSoft Common Objects in COntext)[12]的平均精度(Average Precision，AP)比基线模型提升1.8个百分点，在更具挑战性的Butterfly_2018和Butterfly_2023数据集的AP值比基线模型分别提升1.3个百分点和0.8个百分点. 图1(a)的AI-DETR模型(虚线)与基线模型Conditional DETR(实线)在MS-COCO验证集的AP值(纵坐标)随训练轮次(横坐标)的变化曲线表明：新模型AI-DETR的性能优于基线模型. 图1(b)~(d)的AI-DETR模型与基线模型在MS-COCO训练集的总损失、预测框回归损失、交叉熵分类损失随训练轮次的变化曲线(横坐标为训练轮次，纵坐标为损失值)揭示：新模型AI-DETR的收敛性优于基线模型. 文章附录G展示了AI-DETR模型的更多实验结果，有效验证了本文方法在提升模型收敛性、检测精度和普适性方面的优越性.

2 相关工作

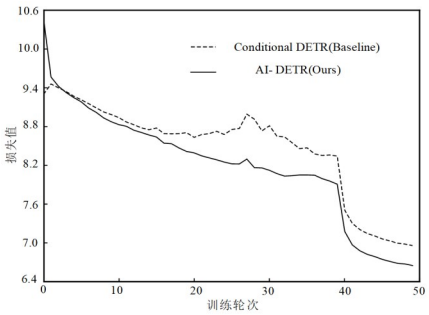
本节介绍与本文相关的研究工作，包括基于DETR的系列目标检测算法及基于深度学习的蝴蝶检测研究.

2.1 基于DETR的系列目标检测算法

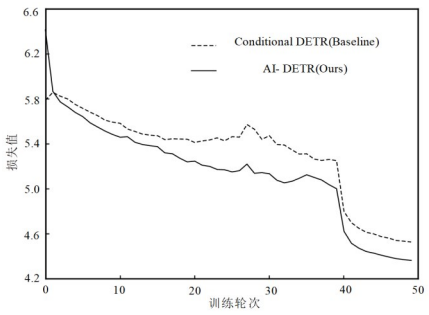
DETR[5]是首个将Transformer应用于目标检测的方法，将目标检测转化为集合预测问题，摒弃基于卷积神经网络的检测算法所需的锚框机制和非极大值抑制后处理，实现真正意义上的端到端检测. 然而，DETR存在小物体识别精度差和收敛速度慢等问题. 因此，不少研究者提出DETR的改进方案. Deformable DETR[7]将可变形卷积稀疏采样与Transformer全局关系建模相结合，提出可变形点采样注意力机制，降低模型计算量，在此基础上，使用多尺度特征提升小目标检测性能. Conditional DETR[8]将交叉注意力空间和语义计算解耦，避免语义和空间信息互相干扰，引入空间条件查询集，为查询集提供参考点约束，有效提升了模型收敛速度. 动态锚框检测变换器(Dynamic Anchor Boxes DETR，DAB-DETR)[9]在Conditional DETR二维空间嵌入编码基础上引入尺度信息，将其拓展至四维位置先验，并在解码器中逐层校正位置编码，指导注意力关注目标可能存在的区域，进一步提高模型收敛速度和精度. 去噪检测变换器(DeNoising DETR，DN-DETR)[11]基于去噪训练思想解决DETR二分图匹配不稳定问题，在真实目标中添加微小噪声扰动，采用去噪训练方法直接重建目标，辅助模型收敛. 改进的去噪锚框检测变换器(DETR with Improved denoising anchor boxes，DINO)[10]在Deformable DETR[7]、DAB-DETR[9]、DN-DETR[7,9,11]基础上，兼顾精度和收敛速度，提出了基于对比学习的去噪训练、混合查询集选择和预测框迭代优化三种改进策略，使DETR类型方法首次在目标检测领域取得最先进技术



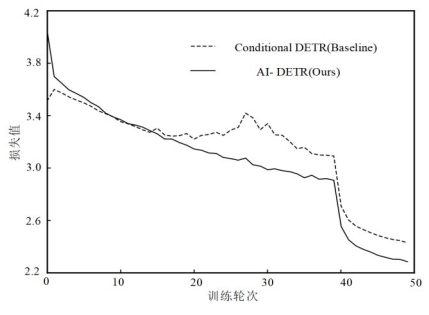
（*a*）平均精度随训练轮次变化曲线



（*b*）总损失随训练轮次变化曲线



（*c*）预测框回归损失随训练轮次变化曲线



（*d*）分类损失随训练轮次变化曲线

图1　本文AI-DETR模型与基线模型Conditional DETR在MS-COCO的各类指标变化曲线

(State-Of-The-Art,SOTA)性能. Decoupled DETR[13]根据分类和回归任务不同,提出了任务感知查询集生成模块,将解码器交叉注意力计算解耦,优化特征学习过程. 混合匹配检测变换器(DETRs with Hybrid matching,H-DETR)[14]提出集合混合匹配方案,在训练过程中使用两个分支分别进行一对一和一对多匹配,当测试时,仅使用一对一匹配分支,显著提升集合匹配准确性. 协作混合分配训练的检测变换器(DETRs with Collaborative hybrid assignments training,Co-DETR)[15]提出协作混合分配训练方案,使用通用的一对多标签分配提高编码器训练效率和解码器训练有效性.

尽管不同的改进策略显著提升了DETR的收敛速度和检测精度,但改进方法多集中在交叉注意力机制中查询集优化、计算方法改进或匹配算法优化等方面,引入模块或参数较多. 为此,本文聚焦解耦交叉注意力机制计算过程,在保持原有注意力机制计算方法情况下,引入非常少参数,增强模型性能.

2.2　基于深度学习的蝴蝶目标检测研究

蝴蝶与环境密切相关,是国际公认的环境和生物多样性监测指示物种,而蝴蝶识别常需要专家参与,耗时耗力,不少研究者借助计算机视觉技术实现蝴蝶物种自动识别. 文献[16,17]为第三届中国数据挖掘竞赛暨国际首次蝴蝶识别大赛提供了一个包含蝴蝶标本图像和野外环境下蝴蝶图像(即蝴蝶生态图像)的数据集,并采用当时最新的Faster RCNN[2]算法实现蝴蝶自动检测,作为竞赛基线模型. 文献[18]使用蝴蝶图像数据集[16,17],提出了基于迁移学习和可变形卷积的蝴蝶检测算法,实现了生态照蝴蝶检测. 针对蝴蝶图像数据集[16,17]样本的长尾分布,文献[19]提出了一种新的数据增强和数据集划分技术,并通过大量实验发现RetinaNet[20]是当时最好的蝴蝶目标检测模型. 在此研究基础上,文献[21]通过统计数据集中图像尺寸、实例宽高比等特征,基于通道注意力机制改进RetinaNet,引入可变形卷积增强网络特征提取能力,通过结果可视化充分解释了改进模型的有效性,并发现影响模型性能的是生态照片中同类蝴蝶图像的结构不一致. 文献[22]提出了具有分布感知惩罚机制的对等学习网络,用于学习图像细粒度特征表示,有效缓解长尾分布数据集存在的类间偏差过大问题. 文献[23]基于人类视觉感受野机制,提出了针对野外环境下蝴蝶识别的改进的感受野模块结合K均值(K-means)聚类和软非极大值抑制网络KSRFB-net(improved Receptive Field Block combined with K-means and Soft non-maximum suppression)新模型,模型中提出通过减小原始感受野模块的感受野来改进的感受野模块(impoved Receptive Field Block by reducing the receptive field of original RFB,RFB-r)模块增强感受野模块(Receptive Field Block,RFB)特征学习能力,采用K-means选择最合适的默认框,并用软非极大值抑制(Soft Non-Maximum Suppression,Soft-NMS)解决一幅图像的多只蝴蝶同时检测问题. 在此工作基础上,文献[24]提出不同空洞率和离心率的多感受野

模块(Multi-Receptive Field Block,MRFB)模块,联合提取不同大小目标的特征,提升模型对野外环境下不同尺寸蝴蝶的检测性能. 文献[25]提出了基于显著性特征的弱监督蝴蝶检测模型,解决蝴蝶图像数据精细标注成本高昂问题. 文献[26]提出了深度学习模型不可知元学习(Deep Learning Model-Agnostic Meta-Learning,DL-MAML)模型,采用元学习解决现有野外环境下蝴蝶物种识别面临的泛化难问题.

基于生态图像的蝴蝶自动识别给目标检测方法带来巨大挑战,研究结果缺少可解释性、模型性能有待提升,提出的自适应可解释模型AI-DETR将突破现有研究局限,并在该富有挑战性的数据集验证其性能.

3 本文方法

本节首先对DETR模型进行简要介绍,然后详细论述本文各个创新点,包括将交叉注意力机制计算解耦,经过定性和定量分析,提出了层间自适应注意力权重更新方法、值向量自适应增强方法,分别从优化注意力权重计算过程和语义增强两个角度提升交叉注意力信息聚合性能,以及为保证预测参考点选择灵活性,提出IRPR方法.

3.1 方法概述

DETR模型主要包括卷积神经网络backbone、Transformer架构的编码-解码器,以及负责分类和边界框回归的预测头3部分. 经过卷积神经网络,模型获得图像局部特征,降低DETR模型整体计算量. Transformer编码器利用自注意力机制将局部特征转为全局特征,解码器使用自注意力机制优化查询集,使用交叉注意力机制聚合查询集与全局特征. 预测头网络基于集合匹配思想从DETR模型解码器输出特征中得到预测结果. DETR模型交叉注意力机制存在语义与空间信息冲突,查询集与全局特征聚合过程中无空间约束,搜索范围无引导信息,影响模型收敛速度. Conditional DETR[8]使用二维参考点约束查询集搜索空间,改良DETR交叉注意力计算策略,该策略被后续DAB-DETR[9]和DINO[10]沿用. 本文将Conditional DETR[8]作为基线模型,对其进行进一步改进.

DETR解码器中交叉注意力机制是模型推理的关键,往往采用多头计算策略,捕捉图像不同区域特征. 每个注意力头计算方法如式(1)所示,其中$\boldsymbol{Q}$表示查询目标向量,通常随机初始化,$\boldsymbol{K}$表示键向量,由DETR模型编码器最后1层输出特征与位置编码组成,$\boldsymbol{V}$表示值向量,是DETR模型编码器最后1层的输出特征,d_k表示$\boldsymbol{K}$的属性维度,Softmax(*)为归一化指数函数,用于生成注意力权重. 因此,式(1)可解耦为注意力权重和值向量2部分.

$$\text{attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Soft max}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

基于解耦后的交叉注意力机制,提出可解释的改进方法,针对基线模型Conditional DETR参考点缺乏灵活性问题,提出IRPR方法提升模型性能. 针对解码器学习偏差的层间传递,提出IAAWR方法,动态调节解码器不同层的交叉注意力权重,削弱偏差的层间传递. 针对编码器输出特征语义模糊,提出值向量AFE方法,改善编码器的特征提取能力,增强输出特征语义. 本文改进模型的整体框架如图2所示,其中只涉及基线模型中Transformer编码器和解码器的简化结构,即本文方法所需部分,实线箭头表示前向传递,虚线箭头表示对箭头指向变量有贡献,α和β分别是注意力权重和值向量更新系数,其下标为层索引,全连接(Fully Connected,FC)为参考点生成网络,解码器各层共享FC参数.

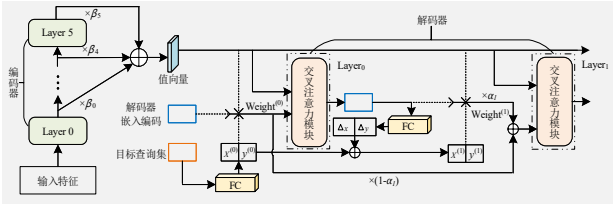


图2 本文AI-DETR模型的整体框架

3.2 层间自适应注意力权重更新

Conditional DETR解码器连续堆叠6层结构一致的解码模块,各层共享预测头参数,某层损失函数值越大,说明该层输出特征偏差越大,反之亦然. Conditional DETR开源代码提供了模型训练日志,图3展示了以ResNet50为Backbone的Conditional DETR模型解码器不同层的训练损失变化曲线. 横坐标为训练轮次,纵坐标为损失值,图例train_loss_i($i\in\{0,1,2,3,4,5\}$)表示第i层的损失.

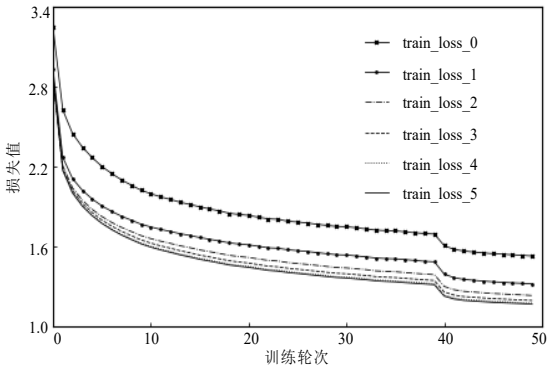


图3 Backbone为ResNet50的Conditional DETR的解码器不同层的训练损失曲线

由图3所示的Conditional DETR模型解码器不同层的训练损失曲线可见，该解码器第1层输出的训练损失值（train_loss_0曲线）远高于其余层，这说明第1层输出与期望输出存在较大偏差. 解码器层层堆叠，各层输出偏差会层层传递，影响模型收敛. 尽管如此，图3结果还显示，各层输出偏差逐层递减，第5、6层输出偏差（train_loss_4和train_loss_5曲线）几乎重合，且是各层输出偏差中最小的.

在附录E图E1所示的Conditional DETR的开源代码中，backbone为其他网络类型的损失函数日志，也显示其解码器不同层的损失函数曲线有同样趋势. 结合3.1节，交叉注意力机制解耦后，注意力权重受偏差直接影响，由此我们分析解码器不同层间交叉注意力权重的关系，定量证明解码器输出偏差存在层层传递现象，具体细节见附录A.

上述定性和定量分析说明，基线模型Conditional DETR的解码器中，不同层间交叉注意力权重正相关，第1层学习偏差存在层层传递问题，增加了训练难度. 因此，本文提出IAAWR方法，使模型动态避免偏差传递. 附录B.1展示了IAAWR方法的6种不同组合方式. 附录B.1的表B2所示消融实验结果表明：IAAWR法的方式I效果最好，其计算过程为

$$\boldsymbol{w}'_{j,k}=\alpha_j\times\boldsymbol{w}_{j,k}+(1-\alpha_j)\times\boldsymbol{w}'_{j-1} \tag{2}$$

式中，$\boldsymbol{w}'_{j,k}$为更新后的第j+1层第k+1注意力头权重；j和k分别为层索引和注意力头索引；α_j为第j+1层可学习权重，且$\alpha_j\in(0,1)$；$\boldsymbol{w}_{j,k}$为第j+1层第k+1注意力头权重，$j\in\{1,2,3,4,5\}$，$k\in\{0,1,2,3,4,5,6,7\}$；$\boldsymbol{w}'_{j-1}$为第j层更新后的注意力头权重均值. $\boldsymbol{w}'_{j-1}$的计算过程为

$$\boldsymbol{w}'_{j-1}=\frac{1}{8}\sum_{k=0}^{7}w'_{j-1,k} \tag{3}$$

结合式（3），将式（2）转换为式（4），即附录B. 1节表B1的方式I，表示相邻层间所有交叉注意力头权重的迭代更新.

$$\boldsymbol{w}'_j=\alpha_j\times\boldsymbol{w}_j+(1-\alpha_j)\times\boldsymbol{w}'_{j-1} \tag{4}$$

Conditional DETR解码器中交叉注意力模块共堆叠6层. 结合式（4），方式I可拓展为任意层所有交叉注意力头权重的迭代更新，计算方式如式（5）所示，J为当前层索引，$J\in\{1,2,3,4,5\}$，$\boldsymbol{w}_0$为第1层所有头交叉注意力权重均值，λ_j取值如式（6）所示，α_j同式（2）. 式（5）和式（6）的推导过程见附录C.

$$\boldsymbol{w}'_J=\sum_{j=1}^{J}\lambda_j\times(\boldsymbol{w}_j-\boldsymbol{w}_0)+\boldsymbol{w}_0 \tag{5}$$

$$\lambda_j=\begin{cases}\alpha_j\times(1-\alpha_{j+1})\times\cdots\times(1-\alpha_J), & 1\leqslant j<J\\ \alpha_j, & j=J\end{cases} \tag{6}$$

由式（5）可知，$\boldsymbol{w}_0$作为其余层注意力权重基底，可为高层注意力权重计算提供依据，同时为避免$\boldsymbol{w}_0$偏差传递，使用λ_j动态调节. 据此，IAAWR方法既可在基底辅助下，减轻注意力权重学习难度，又可自动规避偏差传递.

3.3　值向量自适应增强

式（1）值向量$\boldsymbol{V}$为模型编码器最后1层的输出特征，为交叉注意力提供语义信息. 基于卷积神经网络的目标检测工作[27]指出卷积神经网络浅层关注小目标，深层关注大目标，通过将不同网络层输出特征融合，可增强特征语义信息，提升模型性能.

鉴于卷积和Transformer的计算差异，本文首先对基线模型编码器各层输出特征进行分析，选取MS-COCO验证集中序号分别为147338、435003、571598和581317的图像输入训练好的基线模型，将编码器不同层输出特征图可视化，如图4所示，第1列为每行输入图像，其余列依次为编码器各层对输入图像的特征图逐通道加和求均值后的特征表示.

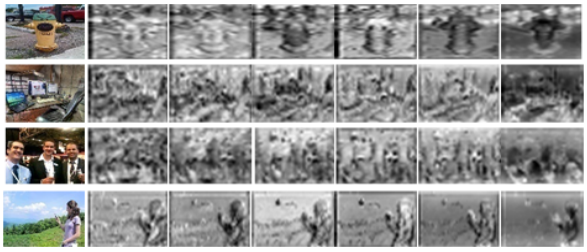


图4　编码器不同层输出特征逐通道加和求均值后的表示

图4的可视化结果定性说明了基线模型编码器各层输出特性：基线模型编码器不同层提取的语义信息有相似，也有差异，尤其是编码器最后1层与其余层特征表示差异明显. 同时，部分样本的特征表示缺乏显著结构信息，例如，图4第2行图像435003在编码器各层的特征表示.

为进一步探明编码器不同层输出特征的关系，我们使用余弦相似性定量度量图4中各样本经模型编码器提取特征后，各相邻层特征表示的区别与联系，结果四舍五入保留4位小数后如表1所示，序偶表示相邻层索引，最后1列为对应行（图像）相似性的均值.

表1　各样本经基线模型编码器后相邻编码层特征相似性

图像序号	<0,1>	<1,2>	<2,3>	<3,4>	<4,5>	相似性均值
147338	0.995 0	0.990 9	0.995 0	0.994 4	–0.639 0	≈0.667 3
435003	0.991 6	0.991 7	0.993 9	0.996 7	–0.712 3	≈0.652 3
571598	0.985 5	0.990 9	0.990 4	0.991 0	0.075 1	≈0.806 6
581317	0.994 9	0.995 4	0.995 2	0.997 1	–0.451 3	≈0.706 3

由表1可知，除最后一层外，编码器其余层特征均值相似性极高，说明编码器存在学习冗余，最后一层与其余各层呈负相关或弱相关，说明最后一层与其余层

输出特征差异大. 这一结果与图4定性显示的结果一致.

综合图4定性分析和表1定量分析可知,基线模型Conditional DETR将编码器最后一层输出作为值向量,存在信息丢失问题. 因此,受文献[27]启发,本文提出了简单有效的值向量AFE方法.

由于编码器各层输出特征图尺寸一致,本文对不同层输出特征直接加权实现特征自适应增强,加权方式如式(7)所示,β_t为第t+1层输出特征权重,$\boldsymbol{v}_t$为第t+1层输出特征,$\boldsymbol{V}'$为增强后的值向量.

$$\boldsymbol{V}'=\sum_{t=0}^{5}\beta_t\times\boldsymbol{v}_t \tag{7}$$

由式(7)可知,$\boldsymbol{V}'$采用分治思想动态组合模型编码器不同层的输出特征,以期达到抑制编码器各层冗余学习,丰富值向量语义的效果.

3.4 IRPR方法

基线模型Conditional DETR使用二维参考点约束交叉注意力计算区域,训练和推理过程中,Conditional DETR解码器不同层的交叉注意力计算使用相同参考点,未进行层间微调,缺乏灵活性. 为此,本文提出IRPR方法对解码器每层坐标进行矫正,具体计算过程如式(8)和(9)所示. 其中,$\boldsymbol{q}_u$表示解码器第u+1层输出特征,ref$\boldsymbol{P}_{u-1}$表示解码器第u层的参考点标准化坐标,ref$\boldsymbol{P}'_u$表示解码器第u+1层参考点非标准化向量,ref$\boldsymbol{P}_u$表示解码器第u+1层参考点标准化坐标. 其中,$u\in\{0,1,2,3,4,5\}$. 为简化计算复杂度,layer$_u$采用基线模型中已有的生成初始参考点的两层全连接网络,层间使用修正线性单元(Rectified Linear Unit,ReLU)函数引入非线性,inverseSigmoid(*)表示Sigmoid(*)逆运算.

$$\text{ref}\,\boldsymbol{P}'_u=\text{layer}_u(\boldsymbol{q}_u)+\text{inverseSigmoid}(\text{ref}\,\boldsymbol{P}_{u-1}) \tag{8}$$

$$\text{ref}\,\boldsymbol{P}_u=\text{Sigmoid}(\text{ref}\,\boldsymbol{P}'_u) \tag{9}$$

由式(8)可知,当IRPR方法在对解码器每层坐标进行矫正时,在先前层参考点坐标基础上,充分考虑了当前层输出特征的语义信息. 整个过程不引入额外参数. 为稳定更新参考点,网络训练过程中对解码器第u+1层参考点计算梯度时,不考虑ref$\boldsymbol{P}_{u-1}$对梯度的影响.

4 实验结果与分析

本节将介绍实验设计细节和实验结果,并对实验结果进行详细分析. 其中,4.1节阐明实验设计,4.2节通过实验验证本文选用的基线模型的合理与正确性,4.3节测试本文提出方法的通用性,4.4节通过消融实验验证本文所提方法的有效性,4.5节展示本文提出的AI-DETR模型在极具挑战性的野外环境下蝴蝶数据集实验结果,进一步验证本文提出方法的强大功能.

4.1 实验设计

实验首先使用公开数据MS-COCO[12]验证本文提出方法的有效性,接着使用离线扩增的第三届中国数据挖掘竞赛暨国际首次蝴蝶识别大赛数据集Butterfly_2018[16,17,21]和陕西师范大学计算机科学学院实验室私有的数据集Butterfly_2023,验证本文所提方法在细粒度非平衡数据集的性能. 上述数据集的基本信息如表2所示. 蝴蝶数据集的更进一步信息见附录D.

表2 实验数据集基本信息描述

数据集名称	训练集图像数量	测试集图像数量	类别数
MS-COCO	118 287	5 000	80
Butterfly_2018	12 070	687	94
Butterfly_2023	19 601	2 127	324

实验基于Conditional DETR开源代码. 当使用MS-COCO数据集[12]进行实验时,保持源代码设置. 针对Butterfly_2018和Butterfly_2023数据集,为加快推理速度和节省计算资源,在输入图像保留原始长宽比情况下,调整为最长边不超过800像素,其余设置与源代码保持一致. 实验结果使用MS-COCO评价指标[12],所有结果用百分数表示,AP表示所有阈值下平均精度的均值,AP_{50}、AP_{75}、AP_{85}和AP_{95}分别表示在阈值为0.5、0.75、0.85和0.95时的平均精度. 阈值越高,预测框越准确. AP_S、AP_M和AP_L分别为小、中和大目标的平均精度,AR_{10}表示每张图片预测框数为10的平均查全率. 实验中对比的其他方法均使用官方开源代码. 因实验资源有限,程序的训练和推理结合模型所需显存切换显卡,同一模型均采用单张相同型号显卡. 具体是Conditional DETR和DAB-DETR使用Nvidia GeForce RTX 3080Ti,DINO-*使用Nvidia GeForce RTX 3090.

4.2 基线模型选择实验

在DETR系列工作中,Conditional DETR、DAB-DETR和DINO是引入动态查询集嵌入编码机制的代表工作,本文对这三种方法进行比较,选出合适的基线模型. 为保证可比性,各模型均采用ResNet50[28]作为backbone. 为避免硬件差异,各模型均在官方开源预训练模型基础上,在本地环境下使用MS-COCO训练集进行微调,各模型在MS-COCO验证集的实验结果如表3所示. 表3中GFLOPs(Giga FLoating-point Operations)为10亿次浮点运算次数,每秒传输帧数(Frames Per Second,FPS)为模型每秒能推理的图像数量,其中不带"*"为原论文实验结果,带"*"为本地实验环境下微调后实验结果,"—"表示原论文未提及,括号中值反映较原论文实验结果AP的增减情况,DINO-4s和DINO-5s分别表示使用ResNet50输出的4个和5个尺度特征图,最佳结果用加粗表示.

表3实验结果表明:DINO-4s模型由于加入了多种

表3　本地环境下各模型在MS-COCO验证集的实验结果

模型名称	AP/%	AP_{50}/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%	GFLOPs /G	FPS/s
Conditional DETR	40.9	61.8	43.3	20.8	44.6	59.2	90	—
DAB-DETR	42.2	63.1	44.7	21.5	45.7	60.3	94	—
DINO-4s	49.0	66.6	53.5	32.0	52.3	63.0	279	24
DINO-5s	49.4	66.9	53.8	32.3	52.5	63.9	860	10
Conditional DETR*	40.1(−0.8)	60.2	42.4	19.6	43.5	58.7	90	14
DAB-DETR*	40.1(−2.1)	60.2	42.4	17.5	43.8	61.5	94	14
DINO-5s*	49.7(+0.5)	66.7	54.2	**33.4**	52.5	64.0	860	9
DINO-4s*	**50.4(+1.4)**	**67.9**	**55.1**	32.5	**54.0**	**64.3**	279	16

优化策略，在本地实验环境下微调时得到小幅提升，除AP_S外，AP其余各项指标达到最佳，但计算量是其他两个模型的近3倍. Conditional DETR和DAB-DETR在本地实验环境下，微调后的实验结果均较原文性能有所下降，但Conditional DETR下降幅度更小，对硬件敏感性低且推理速度与DINO-4s*接近，但计算量最小. 因此，本文选择Conditional DETR作为基线模型.

4.3　通用性测试实验

为验证本文提出方法的通用性，本小节采用两类实验进行验证. 首先，将本文提出的方法嵌入不同模型进行实验；其次，将本文提出的所有方法嵌入不同Backbone的Conditional DETR基线模型进行实验. 以此验证提出方法的通用性. 因此，首先用所提方法增强基线模型，并对DAB-DETR和DINO模型嵌入所提方法. 其中，Conditional DETR和DAB-DETR加入本文提出的所有方法，DINO模型由于采用可变形注意力机制和预测框细化方法，因此，在DINO中仅引入提出的值向量AFE方法，各模型backbone均为ResNet50. 各改进后模型在MS-COCO验证集的结果如表4所示，“#”表示加入本文方法改进的模型，括号内值表示与表3的“*”模型相比，各模型的指标增减情况，最佳结果用加粗表示.

接着，考虑基线模型Conditional DETR使用了4种backbone，其中低分辨率backbone为RestNet50和ResNet101，输出32倍下采样特征图，记为R50和R101；高分辨率backbone为DC5-ResNet50和DC5-ResNet101，输出16倍下采样特征图，记为DC5-R50和DC5-R101，DC5表示ResNet最后1个阶段采用空洞卷积，在不进行下采样基础上扩大感受野. 使用本文提出的所有方法改进4种不同backbone的基线模型，在MS-COCO验证集进行测试，实验结果如表5所示，表中“#”表示本文提出的所有方法改进后模型，括号内值反映较原模型对应指标的增减情况.

表4　不同模型嵌入本文所提方法后在MS-COCO验证集的测试结果

模型名称	AP/%	AP_{50}/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%	FPS/s
Conditional DETR$^{\#}$	41.9(+1.8)	62.2(+2.0)	44.3(+1.9)	21.5(+1.9)	45.1(+1.6)	60.4(+1.7)	13(−1)
DAB-DETR$^{\#}$	41.3(+1.2)	61.3(+1.1)	43.8(+1.4)	17.6(+0.1)	45.0(+1.2)	63.2(+1.7)	13(−1)
DINO-5s$^{\#}$	50.4(+0.7)	67.4(+0.7)	54.8(+0.6)	33.6(+0.2)	53.3(+0.8)	63.6(−0.4)	9(0)
DINO-4s$^{\#}$	**50.7**(+0.3)	**68.4**(+0.5)	**55.2**(+0.1)	**33.8**(+1.3)	**53.7**(−0.5)	**65.1**(+0.8)	15(−1)

由表4可知，各模型的AP指标在嵌入本文所提方法后均有提升，FPS与原模型基本相当，说明本文提出的方法在保证推理效率的前提下，对各模型整体性能均有正向作用. 表4中各模型的其余评价指标，除DINO-5s的AP_L和DINO-4s的AP_M指标有少许降低外，各改进模型的性能都有提升. 由于本文所提方法引入的额外参数几乎可忽略不计，当模型参数规模增大时，计算复杂性变高，少量参数带来的性能增强会逐步减弱. 表4改进模型Conditional DETR$^{\#}$的AP增益最高，然后依次是DAB-DETR$^{\#}$、DINO-5s$^{\#}$和DINO-4s$^{\#}$的AP增益，该结果也验证了随模型规模变大，本文所提方法带来的模型性能增益逐渐降低.

表5实验结果表明：在不同backbone的基线模型中嵌入本文提出的所有方法，各模型的所有评价指标均有明显提升，FPS与原模型相当，说明本文提出的方法对高、低分辨率的backbone均有效. 此外，当使用本文所提方法改进不同backbone的基线模型时，模型增益对模型backbone输出特征图大小敏感. 例如，R50$^{\#}$模型和R101$^{\#}$模型的backbone均输出32倍下采样的特征图，而DC5-R50$^{\#}$模型和DC5-R101$^{\#}$模型的backbone均输出16倍下采样的特征图，前两者的性能增益明显强于后两者. 特征图尺度越小，目标的表征越抽象，同时模型整体的计算量越小. 因此，本文提出的方法对抽象的目标特征更有效. 基线模型DC5-R50的AP值高于R50模型，而改进后的DC5-R50$^{\#}$的AP弱于改进后的R50$^{\#}$，进一步验证了本文提出方法倾向于显著增强抽象特征.

表5实验结果还显示，无论基线模型还是改进模型，backbone参数规模越大，模型性能越强，如R101$^{\#}$性

表5 所提方法改进的不同骨干网络的基线模型在MS-COCO验证集的实验结果

骨干网络	AP/%	AP_{50}/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%	FPS/s
R50	40.1	60.2	42.4	19.6	43.5	58.7	14
R101	40.2	60.3	42.8	19.8	44.0	57.8	12
DC5-R50	40.6	60.5	42.6	18.6	44.5	60.8	15
DC5-R101	41.0	60.9	43.5	18.8	45.3	61.4	15
R50#	41.9(+1.8)	62.2(+2.0)	44.3(+1.9)	21.5(+1.9)	45.1(+1.6)	60.4(+1.7)	13(−1)
R101#	42.0(+1.8)	62.0(+1.7)	44.7(+1.9)	22.5(+2.7)	45.7(+1.7)	60.1(+2.2)	12(0)
DC5-R50#	41.6(+1.0)	61.5(+1.0)	44.0(+1.4)	19.7(+1.1)	45.9(+1.4)	61.8(+1.0)	15(0)
DC5-R101#	42.2(+1.2)	62.0(+1.1)	45.0(+1.5)	19.6(+0.8)	46.5(+1.2)	61.8(+0.4)	14(−1)

能强于R50#，DC5-R101#性能强于DC5-R50#，说明目标局部表征影响模型整体性能. 为减少训练时间，本文后续实验和可视化采用R50模型，即backbone为ResNet50的基线模型Conditional DETR.

表4将本文方法嵌入不同模型的实验结果，以及表5将本文方法嵌入不同backbone的Conditional DETR基线模型的实验结果，分别验证了本文所提方法对不同模型的通用性和有效性，以及对不同backbone的Conditional DETR模型的正向兼容性. 附录G展示了本文提出方法的更多实验结果.

4.4 消融实验

本节通过消融实验，验证本文各个创新点IAAWR、值向量AFE和IRPR方法对模型性能的贡献. 表6为本文提出的IAAWR、值向量AFE和IRPR方法在MS-COCO验证集的消融实验结果，“√”表示在基线模型R50（Backbone为ResNet50的Conditional DETR）中加入对应方法，“×”表示未加入，最佳结果用加粗表示.

表6 本文提出的IAAWR、值向量AFE和IRPR方法改进基线模型R50(Conditional DETR)的消融实验结果 单位:%

IAAWR	AFE	IRPR	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
×	×	×	40.1	60.2	42.4	19.6	43.5	58.7
√	×	×	40.8	61.0	43.5	19.8	44.5	59.3
×	√	×	41.1	61.4	43.4	21.0	44.9	58.9
×	×	√	40.4	60.5	43.1	19.8	43.6	58.3
√	√	×	41.5	61.7	44.2	21.0	**45.1**	60.1
√	×	√	40.9	60.9	43.7	20.5	44.4	58.9
×	√	√	40.7	60.8	43.3	20.9	44.1	59.1
√	√	√	**41.9**	**62.2**	**44.3**	**21.5**	**45.1**	**60.4**

表6消融实验结果表明：在增加单组件情况下，值向量AFE方法对基线模型性能提升最高，仅引入6个可学习参数就将基线模型的AP提升1个百分点，说明优化语义信息对交叉注意力的重要性. 在增加两个组件情况下，IAAWR和值向量AFE组合方法对基线模型性能提升最高，在仅引入11个可学习参数情况下，使基线模型的AP增加1.4个百分点. 本文所提三种组件联合使基线模型性能提升最多，AP提升1.8个百分点，且模型在各指标均达到最佳结果. 本文提出的各组件的详细消融实验见附录B.

4.5 蝴蝶数据集实验

为进一步验证本文提出方法的鲁棒性，本小节将提出的IAAWR、值向量AFE和IRPR改进后的R50模型AI-DETR方法用于野外环境下的蝴蝶检测，数据集为Butterfly_2018[21]和Butterfly_2023，这2个数据集均为细粒度非平衡数据集，较MS-COCO数据集更复杂，数据集具体细节见附录D. 表7展示了基线模型R50（Backbone为ResNet50的Conditional DETR）和采用本文AI-DETR模型在这2个蝴蝶数据集的实验结果，最佳结果用加粗表示. 这2个蝴蝶数据集的更多实验结果见附录G的表G2.

表7 基线模型R50和本文AI-DETR在蝴蝶数据集实验结果

单位:%

数据集	模型	AP	AP_{50}	AP_{75}	AP_{85}	AP_{95}	AR_{10}
Butterfly_2018	R50 (Conditional DETR)	63.7	**82.1**	**72.2**	51.4	15.1	72.9
	AI-DETR (Ours)	**65.0**	82.0	71.3	**56.8**	**16.7**	**75.5**
Butterfly_2023	R50 (Conditional DETR)	78.9	89.2	85.7	78.4	33.5	87.0
	AI-DETR (Ours)	**79.7**	**89.7**	**86.2**	**79.2**	**35.3**	**88.3**

表7实验结果表明：无论R50还是嵌入本文提出方法改进的新R50模型AI-DETR，在Butterfly_2023数据集的实验结果均比在Butterfly_2018数据集表现更优，拥有更高的识别精度和平均查全率. 该结果再次验证了本文前期的研究[19]结果：在训练集中，蝴蝶标本图像对提升模型在野外环境下的蝴蝶检测性能无益.

表7实验结果还表明：本文方法改进的R50模型AI-DETR分别在Butterfly_2018和Butterfly_2023数据集的AP值得到1.3个百分点和0.8个百分点的提升，AR_{10}指标值在这2个数据集上分别提升了2.6个百分点和1.3个百分点. 结合附录D可知，Butterfly_2018待识类数和类间样本不平衡性均低于Butterfly_2023，且Butterfly_2018数据集的训练集和测试集相关性更弱，即But-

terfly_2018的源域和目标域分布偏移更大. 根据表7所示的基线模型和提出的AI-DETR模型在这2个蝴蝶数据集的性能表现可知，类别数多且极端长尾分布是导致AI-DETR模型在Butterfly_2023数据集性能增益低于在Butterfly_2018数据集的潜在因素，而域间分布偏移制约模型在Butterfly_2018数据集中的表现，与本文前期研究[21]中发现的制约模型性能的分析结果一致.

表7实验结果还显示，AI-DETR模型在这2个数据集中的整体性能均有提升，说明本文所提方法在真实环境下的复杂数据集上有效. 相对基线模型，本文AI-DETR模型在评价指标AP_{85}和AP_{95}上提升明显，说明本文所提方法能够预测更准确的目标框；AI-DETR模型在AR_{10}指标的提升，说明本文所提方法有效增强图像中目标的发现能力，即同样的输入图像，AI-DETR模型比基线模型检测出更多正确目标. AI-DETR模型在Butterfly_2018数据集的AP_{50}和AP_{75}指标比基线模型R50稍有降低，再次说明训练集中更多类别的标本图像对模型准确检测识别野外环境下的蝴蝶有负面干扰.

5　各创新点性能可视化与可解释性分析

本节将通过展示提出的AI-DETR模型在推理过程中的解码器层间交叉注意力权重的相似性、编码器输出特征图和注意力权重及参考点在原图的映射，为本文AI-DETR模型提出的3个新方法IAAWR、值向量AFE及IRPR方法的性能提供可视化与可解释性分析，部分样本检测结果见附录F.

5.1　解码器层间交叉注意力权重相似性

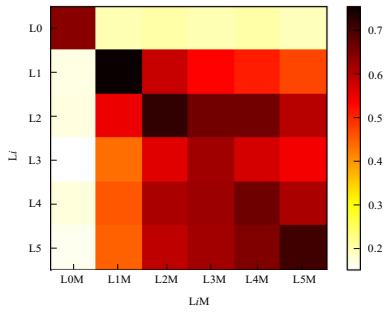
注意力权重通常具有稀疏性，本文采用余弦相似性分析，研究了提出的AI-DETR模型中解码器不同层的交叉注意力权重，分别从全局和局部角度解释AI-DETR模型中的IAAWR方法（即附录B.1中IAAWR方法的方式I)的性能. 从MS-COCO验证集随机选取4张序号分别为147338、435003、571598和581317的图像，使用AI-DERTR模型进行推理，可视化解码器各层每个注意力头权重与各层所有注意力头权重均值相似性的均值，结果如图5(a)~(d)所示，横坐标Li$M(i\in\{0,1,2,3,4,5\})$表示第i层各个注意力头权重均值索引，纵坐标Li($i\in\{0,1,2,3,4,5\}$)表示第i层索引，各图右侧为图例，颜色越深表示相似性越大. 这里的相似性计算采用附录A的式(A1)，从全局角度展示AI-DETR模型解码器各层间交叉注意力权重的相似性，解释其各层间的偏差传递情况. 图6(a)~(d)可视化AI-DETR模型解码器各层各个注意力头权重相似性，横坐标为AI-DETR模型解码器层索引序偶，如25表示第3层和第6层，纵坐标0~7表示注意力头索引，纵坐标8索引行表示相应序偶层8对注意力头权重相似性均值，每幅图右侧为图例，颜色越深表示相似性越强. 这里需要注意的是图6除索引行8外，每行表示不同序偶层对应同一位置的注意力头权重相似性，每列表示相应序偶层各对同一位置的注意力头权重相似性，计算方式如附录A的式(A2)所示. 图6从局部视角定量解释AI-DETR模型解码器各层误差传递情况. 图7展示了AI-DETR模型用的IAAWR方法的5个可学习权重在训练过程中的变化情况，横坐标为训练轮次，纵坐标为权重值.

结合图例，对比图5和附录A的图A1可见，不同样本经过AI-DETR推理后，在全局注意力权重相似性分布方面保持了与基线模型Conditional DETR推理的一致性，即解码器各层各注意力头权重与各层所有注意力头权重均值相似性的均值保持与基线模型大致相同，层自身注意力权重相似性高于与其他层的相似性，第5、6层的相似性最强. 但不同的是，本文模型AI-DETR推理过程中，解码器第1层与其余层的注意力权重均值相似性明显减弱，除第1层外其余层的层间注意力权重相似性明显增强. 这说明AI-DETR模型削弱了解码器第1层偏差的传递，增强了其余层间的信息传递. 因此，AI-DETR模型的IAAWR方法是非常有效的.

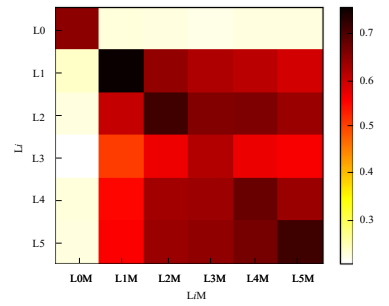
对比图6和附录A的图A2可知，当不同输入样本在AI-DETR模型中推理时，解码器不同层各对注意力头权重相似性存在相同与不同的情况，说明AI-DETR模型与基线模型相同，可提取各样本相同与可辨识的特征. 另外，图6和附录A图A2结果对比显示，AI-DETR模型显著弱化了解码器第1层与其余层注意力权重相似性，强化了除第1层外其余各层的注意力权重相似性，这点与图5结果一致.

以上关于图5、图6的结果分析表明：本文AI-DETR模型的解码器层间注意力权重相似性，无论从全局角度还是从局部角度均显示，解码器第1层与其余层交叉注意力权重之间关联性降低，而其余层间的交叉注意力权重关联性增强. 这说明AI-DETR模型提出的解码器IAAWR方法有效解决了基线模型Conditional DETR的解码器学习偏差的层间传递问题. 结合图1所示的AI-DETR模型有更快的收敛速度和更好的性能表现，进一步说明了本文提出的AI-DETR模型一定程度上缓解了基线模型解码器学习偏差的潜在传递.

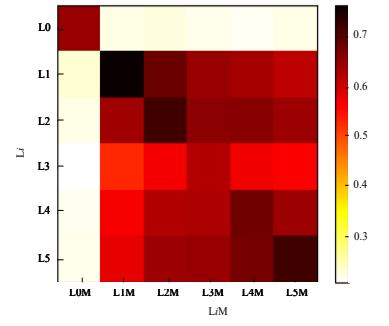
图7展示了本文提出的IAAWR方法（即附录B.1中IAAWR方法的方式I)的5个可学习参数在AI-DETR模型训练过程中的变化情况，结果表明：IAAWR方法的5个可学习参数在模型训练40轮左右明显收敛，5个可学习参数的最终收敛值均接近1. 前40轮各参数由0逐渐增长，在前10轮增长较快，增长速度从快到慢依次分别是α_2、α_4、α_3、α_1和α_5. 其中，α_1为解码器前两层交叉注意力权重间的系数，由于第1层学习偏差的存在，使α_1收



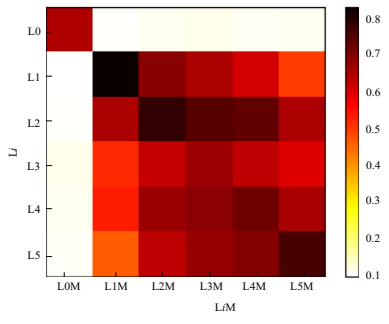
(*a*) 147338



(*b*) 435003

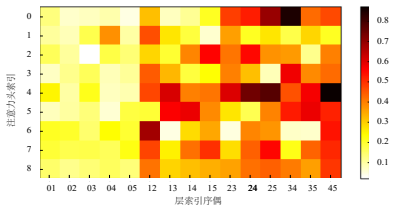


(*c*) 571598

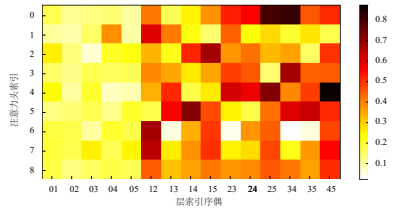


(*d*) 581317

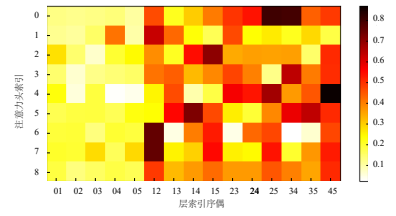
图5 解码器各层各交叉注意力头权重与各层所有交叉注意力头权重均值的相似性均值



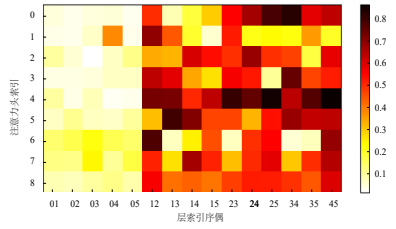
(*a*) 147338



(*b*) 435003



(*c*) 571598



(*d*) 581317

图6 解码器各层各注意力头权重相似性

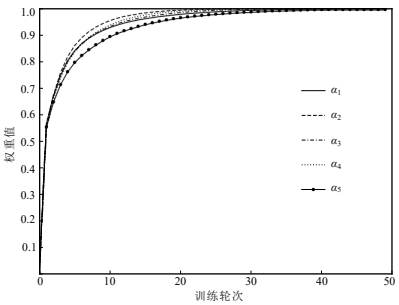


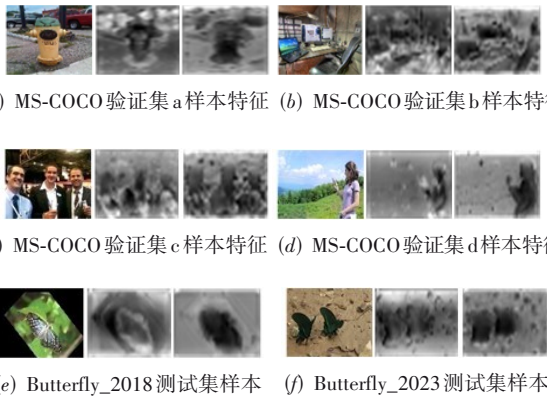
图7 AI-DETR模型使用的IAAWR方法的5个可学习参数在训练过程中随轮次变化情况

敛速度缓慢. α_5为解码器最后两层交叉注意力权重间的系数,由式(5)可见,当最后1层注意力权重更新时,本质上考虑了前面所有层间的联系,导致收敛缓慢.

图7展示的实验结果还表明:IAAWR方法在模型训练过程中,可自动调整解码器第1层的注意力权重对其余各层注意力权重的影响,即训练初期作为其余各层基底,训练后期逐步被其余各层摒弃,有效缓解了潜在的解码器学习偏差传递. 这与式(5)和图5、图6展示的结果一致.

5.2 编码器输出特征图

本节通过可视化基线模型与本文模型编码器层输出的特征图，说明本文模型 AI-DETR 提出的值向量 AFE 方法的有效性．图 8 对比了基线模型 Conditional DETR 和本文模型 AI-DETR 编码器最后一层输出特征逐通道加和求均值后的特征表示，各子图第 1 列为原图，第 2 列为基线模型 Conditional DETR 编码器最后一层输出特征逐通道加和求均值的表示，第 3 列为本文提出的 AI-DETR 模型编码器最后一层输出特征的逐通道加和后均值表示．图 8(a)~(d)为 MS-COCO 验证集样本，图 8(e)为 Butterfly_2018 测试集样本，图 8(f)为 Butterfly_2023 测试集样本．图 9 展示了样本输入模型后，各模型编码器各层输出特征逐通道加和求均值的特征．在图 9(a)~(b)中，左侧为原图，右侧第 1 行为基线模型 Conditional DETR 编码器输出特征的逐通道加和求均值的表示，第 2 行为本文 AI-DETR 模型的编码器输出特征逐通道加和求均值后的表示．模型编码器 6 层分别对应图 9 右侧每行的 6 列．

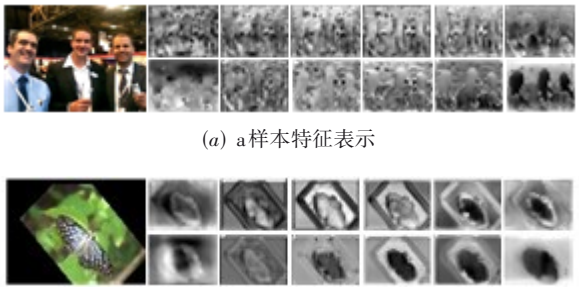


(a) MS-COCO 验证集 a 样本特征　(b) MS-COCO 验证集 b 样本特征

(c) MS-COCO 验证集 c 样本特征　(d) MS-COCO 验证集 d 样本特征

(e) Butterfly_2018 测试集样本　(f) Butterfly_2023 测试集样本

图 8　基线模型 Conditional DETR 和本文 AI-DETR 模型编码器最后一层输出特征逐通道加和求均值后的特征表示



(a) a 样本特征表示

(b) b 样本特征表示

图 9　基线模型 Conditional DETR 和本文 AI-DETR 模型编码器各层各通道输出特征的均值图

由图 8 的可视化结果可知，本文 AI-DETR 模型编码器最后一层输出特征的表示能明显感知到目标的显著性区域．相比之下，基线模型 Conditional DETR 最后一层输出特征表示难以有效区分前景与背景．例如，在图 8(d)中，人手中的手机在基线模型的特征表示中并不明显，而本文 AI-DETR 模型则能够有效感知该目标．由此可见，本文 AI-DETR 模型提出的值向量 AFE 方法有效避免了模型丢失目标的语义信息细节，辅助模型捕捉前景显著性区域，并去除了图像的非目标区域干扰．

由图 9 可视化结果可知，基线模型 Conditional DETR 编码器最后 1 层主要关注图像中的低频部分，即主要语义特征，其余层则关注图像中轮廓或边缘等高频部分．然而，基线模型存在目标特征表示模糊的情况，例如图 9(a)第 1 行基线模型 Conditional DETR 编码器前 5 层输出的特征图，难以区分目标的基本结构．相比之下，本文提出的 AI-DETR 模型的编码器不同层关注图像的不同区域，各层分而治之．编码器每层的特征表示更好地保留了图像结构和细节信息，目标有更显著和准确的边界．例如，图 9(b)第 2 行展示了本文 AI-DETR 模型编码器各层输出的蝴蝶特征表示．此定性分析表明：本文提出的 AI-DETR 模型在特征提取方面具有明显的语义优势．

为了定量描述本文 AI-DETR 模型在提取特征方面的优势，本文将图 9 的样本分别输入本文 AI-DETR 模型和基线模型 Conditional DETR．计算两种模型编码器输出特征逐通道加和求均值后的特征表示，然后将各模型编码器相邻层的特征表示求余弦相似度．结果如表 8 所示，其中序偶表示编码器相邻层索引，最后一列为各模型编码器相邻层特征表示的余弦相似度均值．

表 8　图 9 两样本输入基线模型 Conditional DETR 和本文 AI-DETR 模型后编码器相邻层均值特征的余弦相似度

模型	样本	<0,1>	<1,2>	<2,3>	<3,4>	<4,5>	相似性均值
Conditional DETR	a	0.985 5	0.990 9	0.990 4	0.991 0	0.075 1	≈0.806 6
(Baseline)	b	0.905 7	0.932 1	0.670 2	0.864 7	0.119 7	≈0.698 5
AI-DETR	a	0.824 3	0.838 8	0.852 9	0.940 8	−0.947 8	0.501 8
(Ours)	b	0.340 3	0.687 0	0.896 3	0.928 3	−0.941 4	0.382 1

表 8 结果显示，本文 AI-DETR 模型显著降低了编码器各层输出特征的相似性，有效减少了编码器不同层特征提取的同质性，更加注重分而治之以增强语义表征．具体来看，表 8 中相邻层特征的相似度还显示，本文 AI-DETR 模型明显降低了编码器第 1 层输出特征与第 2 层输出特征的相似性，即序偶层<0,1>特征的相似性．特别是对于样本 b，本文 AI-DETR 模型第 1 层输出特征与第 2 层输出特征的相似性为 0.340 3，相比基线模型

Conditional DETR的0.905 7,降低了0.565 4,约62.4%,大幅减少了第1层特征对后续层学习特征的影响. 对于样本a,本文AI-DETR模型第1层输出特征与第2层输出特征的相似性为0.824 3,相比基线模型Conditional DETR的0.985 5,降低了0.161 2. 然而,表8的结果也表明:本文AI-DETR模型并未全面降低编码器相邻层输出特征的相似度. 例如,对于样本b的序偶层<2,3>和<3,4>,本文模型较基线模型提升了该序偶层输出特征的相似性. 而在序偶层<1,2>和<4,5>,本文模型较基线模型显著削弱了层间输出特征的相似性. 这说明本文模型AI-DETR的编码器具有自适应调整编码器各层输出特征的能力. 对于样本a,本文AI-DETR模型不仅降低了编码器第1层与第2层(即序偶层<0,1>)输出特征的相似性,还降低了其他各相邻层输出特征的相似性. 这再次说明:本文提出的AI-DETR模型的编码器具有自适应地学习编码器各层表示特征的能力,有效解释了AI-DETR模型的值向量AFE方法的有效性.

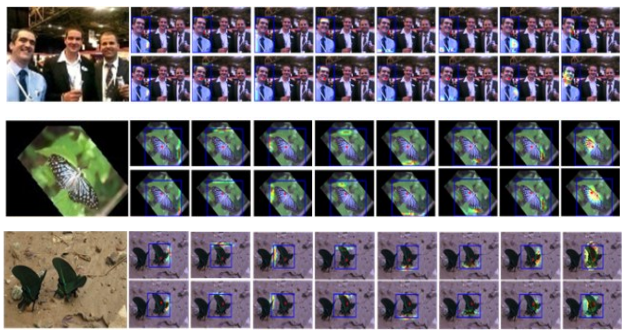
此外,表8中编码器各相邻层均值特征余弦相似度的结果还显示,本文AI-DETR模型与基线模型Conditional DETR最后两层(即序偶层<4,5>)输出特征的相似度明显低于其他相邻层输出特征的相似度. 这是由于编码器最后一层输出特征关注目标语义特征,而其他层输出特征关注目标轮廓等边缘特征. 这一结果与图8的可视化结果的定性分析一致.

综合本小节对AI-DETR模型编码器各层输出特征的定性和定量分析可知,AI-DETR模型不仅显著降低了编码器第1层输出特征对其余层输出特征的影响,还降低了编码器各层特征的提取难度和冗余度,并增强了编码器各层输出特征的语义信息. 这些结果充分证明了AI-DETR模型的值向量AFE方法的有效性.

5.3 解码器注意力权重及参考点在原图的映射

为了进一步展示本文模型AI-DETR提出的IRPR方法的有效性和可解释性,本文从MS-COCO、Butterfly_2018和Butterfly_2023的验证集中各选取1张图像分别送入训练好的基线模型和AI-DETR模型进行推理,筛选分类置信度高于0.9的预测结果,每张图像选择1个预测结果,将推理过程中模型解码器最后1层8个注意力头权重和所用参考点映射到原图,结果如图10所示. 图10A列表示输入图像,每个输入图像对应两行可视化图像,第1行对应基线模型结果,第2行对应本文AI-DETR模型结果,图10B~I列为相应模型解码器最后1层8个注意力头权重在原图的映射,蓝色框为预测框,红色点表示推理所用参考点,暖色调越深表示注意力权重越高,即推理所关注区域.

图10可视化结果表明:无论基线模型还是本文提出的AI-DETR模型,在推理过程中,其解码器8个不同



A B C D E F G H I

注:A表示输入图像;B表示注意力头1权重在原图的映射;C注意力头2权重在原图的映射;D表示注意力头3权重在原图的映射;E表示注意力头4权重在原图的映射;F表示注意力头5权重在原图的映射;G表示注意力头6权重在原图的映射;H表示注意力头7权重在原图的映射;I表示注意力头8权重在原图的映射.

图10 解码器最后一层各注意力头权重及参考点在原图的映射

位置的注意力头关注的区域各不相同. 前7个注意力头倾向关注目标边缘位置,而最后一个注意力头则关注目标的语义特征. 与基线模型相比,本文AI-DETR模型在目标边缘位置的关注区域更加集中,有效减少了目标边缘噪声的影响. 此外,本文AI-DETR模型关注的语义特征区域更准确且广泛,能够更好地获取目标的语义特征.

图10可视化结果还表明:本文AI-DETR模型比基线模型选取的参考点更准确. 例如,在图10的第1个例子中,检测穿浅色衣服的人时,基线模型选取的参考点明显偏离目标所在区域,而本文AI-DETR模型获得的参考点均位于人脸的下巴部位,更准确. 另外,图10第2、3个例子分别来自Butterfly_2018和Butterfly_2023验证集,其推理结果揭示,本文AI-DETR模型选取的参考点比基线模型选取的参考点更偏向于蝴蝶翅膀位置,这更有利于获取蝴蝶分类依据的翅膀特征. 这一结果与人类专家鉴别蝴蝶种类的依据(蝴蝶翅膀背腹面的花纹、颜色与图案)一致. 由此可见,本文AI-DETR模型能依据目标特点动态选择参考点,相较于基线模型具有明显优势. 这些可视化结果充分证明了本文提出的AI-DETR模型的迭IRPR方法的有效性.

6 结论

本文提出了自适应、可解释的目标检测新模型AI-DETR. 该模型提出了解码器IAAWR方法,动态调节解码器第1层粗粒度学习结果对其余层注意力权重的影响,提高了解码器注意力计算稳定性;提出了编码器值向量AFE方法,指导编码器各层分治提取目标区域特征,减少信息丢失,丰富值向量语义;提出了IRPR方法,矫正解码器每层坐标,提升了解码器预测框参考点

灵活性.

AI-DETR模型仅引入了11个可学习参数，却在MS-COCO验证集的AP比基线模型Conditional DETR提升了1.8个百分点，在真实世界的长尾分布数据集Butterfly_2018和Butterfly_2023的AP较基线模型分别提升了1.3个百分点和0.8个百分点. 消融实验还发现，提出的IAAWR和值向量AFE方法联合使基线模型在MS-COCO验证集的AP提升了1.4个百分点；提出的IRPR方法使基线模型在MS-COCO验证集的AP提升了0.3个百分点. 可视化实验结果进一步解释了AI-DETR模型各个创新点的具体贡献，阐述了AI-DETR模型优于基线模型Conditional DETR的内在原因.

同时，本文还选择了性能更强的改进DETR模型DAB-DETR和目前该领域的SOTA模型DINO，融合本文AI-DETR提出的3个创新点，结果使该两个模型的性能均获得提升，再次验证了提出的3个创新点对提升模型性能的通用性. 此外，附录G从本文提出的AI-DETR模型的收敛性、与最新DETR改进模型在MS-COCO的性能对比、与最新蝴蝶检测方法在Butterfly_2018和Butterfly_2023的性能对比，以及AI-DETR模型在自动驾驶、遥感与医学影像等不同场景目标检测任务的性能，进一步验证了提出的AI-DETR模型的优越性，说明本文提出的3个创新点对提升模型性能的有效性.

然而，如何更有效地规避解码器浅层特征偏差传递和增强值向量语义还有待更进一步研究. 另外，本文提出的AI-DETR模型对真实场景下的极端长尾分布数据和域分布偏移数据的鲁棒性也有待进一步提升. 此外，如何有效提升DETR系列模型的训练速度，如采用最近备受关注的Mamba[29,30]架构重构DETR模型，也需要深入研究.

参考文献

[1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 580-587.

[2] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[3] TIAN Z, SHEN C H, CHEN H, et al. FCOS: Fully convolutional one-stage object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 9626-9635.

[4] SUN P Z, ZHANG R F, JIANG Y, et al. Sparse R-CNN: An end-to-end framework for object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15650-15664.

[5] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 213-229.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Neural Information Processing Systems(NIPs). Long Beach: MIT Press, 2017: 5998-6008.

[7] ZHU X, SU W, LU L, et al. Deformable DETR: Deformable transformers for end-to-end object detection[C]//International Conference on Learning Representations(ICLR). Virtual Event: OpenReview.net, 2021: 1-16.

[8] MENG D P, CHEN X K, FAN Z J, et al. Conditional DETR for fast training convergence[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 3631-3640.

[9] LIU S L, LI F, ZHANG H, et al. DAB-DETR: Dynamic anchor boxes are better queries for DETR[C]//International Conference on Learning Representations(ICLR). Virtual Event: OpenReview.net, 2022: 1-20.

[10] ZHANG H, LI F, LIU S L, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection[EB/OL]. (2022-07-11) [2024-11-18]. https://arxiv.org/abs/2203.03605v4.

[11] LI F, ZHANG H, LIU S L, et al. DN-DETR: Accelerate DETR training by introducing query DeNoising[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(4): 2239-2251.

[12] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[M]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[13] ZHANG M Y, SONG G L, LIU Y, et al. Decoupled DETR: Spatially disentangling localization and classification for improved end-to-end object detection[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 6578-6587.

[14] JIA D, YUAN Y H, HE H D, et al. DETRs with hybrid matching[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 19702-19712.

[15] ZONG Z F, SONG G L, LIU Y. DETRs with collaborative hybrid assignments training[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Pis-

cataway: IEEE, 2023: 6725-6735.

[16] 谢娟英, 侯琦, 史颖欢, 等. 蝴蝶种类自动识别研究[J]. 计算机研究与发展, 2018, 55(8): 1609-1618.
XIE J Y, HOU Q, SHI Y H, et al. The automatic identification of butterfly species[J]. Journal of Computer Research and Development, 2018, 55(8): 1609-1618. (in Chinese)

[17] 谢娟英, 曹嘉文, 马丽滨, 等. 蝴蝶物种自动识别研究的生态照片数据集[J]. 中国科学数据, 2019, 4(3): 193-198.
XIE J Y, CAO J W, MA L B, et al. A dataset of butterfly ecological images for automatic species identification[J]. China Scientific Data, 2019, 4(3): 193-198. (in Chinese)

[18] 李策, 张栋, 杜少毅, 等. 一种迁移学习和可变形卷积深度学习的蝴蝶检测算法[J]. 自动化学报, 2019, 45(9): 1772-1782.
LI C, ZHANG D, DU S Y, et al. A butterfly detection algorithm based on transfer learning and deformable convolution deep learning[J]. Acta Automatica Sinica, 2019, 45(9): 1772-1782. (in Chinese)

[19] XIE J Y, LU Y Y, WU Z Z, et al. Investigations of butterfly species identification from images in natural environments[J]. International Journal of Machine Learning and Cybernetics, 2021, 12(8): 2431-2442.

[20] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

[21] 谢娟英, 鲁银圆, 孔维轩, 等. 基于改进RetinaNet的自然环境中蝴蝶种类识别[J]. 计算机研究与发展, 2021, 58(8): 1686-1704.
XIE J Y, LU Y Y, KONG W X, et al. Butterfly species identification from natural environment based on improved RetinaNet[J]. Journal of Computer Research and Development, 2021, 58(8): 1686-1704. (in Chinese)

[22] XU C D, CAI R J, XIE Y H, et al. Fine-grained butterfly recognition via peer learning network with distribution-aware penalty mechanism[J]. Animals, 2022, 12(20): 2884.

[23] XIE J Y, KONG W X, LU Y Y, et al. KSRFB-net: Detecting and identifying butterflies in ecological images based on human visual mechanism[J]. International Journal of Machine Learning and Cybernetics, 2022, 13(10): 3143-3158.

[24] KONG W X, YANG M J, ZHANG J Y, et al. MRFB-net for identifying butterfly species via images taken in the field environments[C]//2023 International Conference on Machine Learning and Cybernetics (ICMLC). Piscataway: IEEE, 2023: 260-267.

[25] ZHANG T, WAQAS M, FANG Y, et al. Weakly-supervised butterfly detection based on saliency map[J]. Pattern Recognition, 2023, 138: 109313.

[26] 赵戈伟, 许升全, 谢娟英. DL-MAML: 一种新的蝴蝶物种自动识别模型[J]. 计算机研究与发展, 2024, 61(3): 674-684.
ZHAO G W, XU S Q, XIE J Y. DL-MAML: An innovative model for automatically identifying butterfly species[J]. Journal of Computer Research and Development, 2024, 61(3): 674-684. (in Chinese)

[27] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 936-944.

[28] 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). Piscataway: IEEE, 2016: 770-778.

[29] GU A, DAO T. Mamba: Linear-time sequence modeling with selective state spaces[EB/OL]. (2024-05-31) [2024-11-18]. https://arxiv.org/abs/2312.00752v2.

[30] ZHU L H, LIAO B C, ZHANG Q, et al. Vision mamba: Efficient visual representation learning with bidirectional state space model[C]//International Conference on Machine Learning (ICML). Vienna: OpenReview.net, 2024: 1-14.

[31] WANG Y M, ZHANG X Y, YANG T, et al. Anchor DETR: Query design for transformer-based detector[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 2567-2575.

[32] GAO P, ZHENG M H, WANG X G, et al. Fast convergence of DETR with spatially modulated co-attention[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 3601-3610.

[33] SUN Z Q, CAO S C, YANG Y M, et al. Rethinking transformer-based set prediction for object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 3591-3600.

[34] ZHAO Y A, LV W Y, XU S L, et al. DETRs beat YOLOs on real-time object detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 16965-16974.

[35] ZHU C C, HE Y H, SAVVIDES M. Feature selective anchor-free module for single-shot object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pat-

tern Recognition (CVPR). IEEE, 2019: 840-849.

[36] ZHANG S F, CHI C, YAO Y Q, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 9759-9768.

[37] ZHOU X Y, WANG D Q, KRÄHENBÜHL P. Objects as points[EB/OL]. (2019-04-25) [2025-03-17]. https://arxiv.org/abs/1904.07850v2.

[38] KONG T, SUN F C, LIU H P, et al. FoveaBox: Beyound anchor-based object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 7389-7398.

[39] ZHU B J, WANG J F, JIANG Z K, et al. AutoAssign: Differentiable label assignment for dense object detection[EB/OL]. (2020-11-25)[2025-03-17]. https://arxiv.org/abs/2007.03496v3.

[40] KIM K, LEE H S. Probabilistic anchor assignment with IoU prediction for object detection[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 355-371.

[41] YANG Z, LIU S H, HU H, et al. RepPoints: Point set representation for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 9657-9666.

[42] CHEN S F, SUN P Z, SONG Y B, et al. DiffusionDet: Diffusion model for object detection[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 19773-19786.

[43] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2012: 3354-3361.

[44] HAROON M, SHAHZAD M, FRAZ M M. Multisized object detection using spaceborne optical imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 3032-3046.

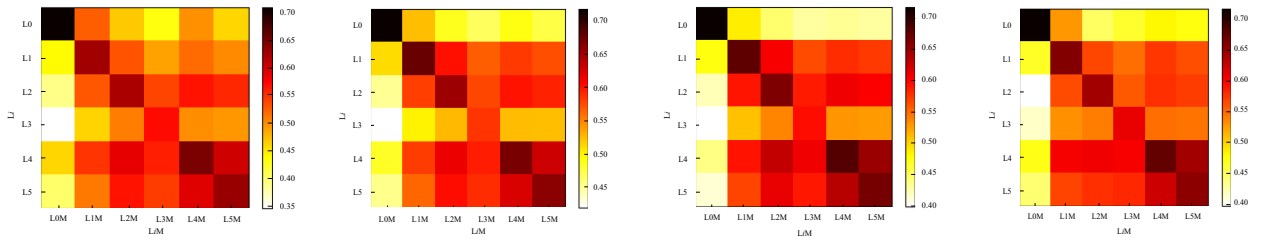
[45] CIAGLIA F, ZUPPICHINI F S, GUERRIE P, et al. Roboflow 100: A rich, multi-domain object detection benchmark[EB/OL]. (2022-12-30) [2025-03-17]. https://arxiv.org/abs/2211.13523v3.

附录A　定量证明解码器输出偏差层层传递

注意力权重通常具有稀疏性，本文通过余弦相似性分析了解码器不同层交叉注意力权重之间的关系. 从MS-COCO[12]验证集中随机选取4张图像，序号分别为147338、435003、571598和581317，使用开源代码公布的、以为ResNet50作为骨干网络(Backbone)的Conditional DETR模型对上述4张图像进行推理. 采用式(A1)计算解码器各层各注意力头权重与各层所有注意力头权重均值的相似性均值，其中，$c, l \in \{0, 1, 2, 3, 4, 5\}$为层索引，$\boldsymbol{w}_{c,k}$为第$c$+1层第$k$+1个注意力头的权重，$\boldsymbol{w}_l$为第$l$+1层所有注意力头权重的均值，similarity(·)为相似性度量函数，本文为余弦相似性函数，h表示每层注意力头数量，k表示头索引，avg(·)为均值函数，$\text{value}_{c,l}$为第c+1层与第l+1层的注意力相似性值. 图A1(*a*)~(*d*)展示了采用式(A1)计算得出的所有相似性值的可视化结果，横坐标表示各层所有注意力头权重均值索引，用L*i*M表示，纵坐标表示层索引，用L*i*表示，其中，$i \in \{0, 1, 2, 3, 4, 5\}$，各图右侧为图例，颜色越深表示相似性越大.

$$\text{value}_{c,l} = \text{avg}\left(\sum_{k=0}^{h-1} \text{similarity}\left(\boldsymbol{w}_{c,k}, \boldsymbol{w}_l\right)\right) \tag{A1}$$

图A1从全局角度说明在Conditional DETR推理过程中，解码器各层间交叉注意力权重具有相似性，不同样本经模型推理后，各层间注意力权重相似性分布大致相同，且各层所有注意力头权重均值与自身层各注意力头权重相似性均值大于与其他层各注意力头权重相似性均值，第1层各注意力头权重与其他各层注意力权重相似性最小，第5、6两层注意力权重相似性最大. 结合注意力权重计算过程说明第1层偏差输出存在层层传递现象，且偏差越来越小，与图3的结果一致.



(*a*) 147338　(*b*) 435003　(*c*) 571598　(*d*) 581317

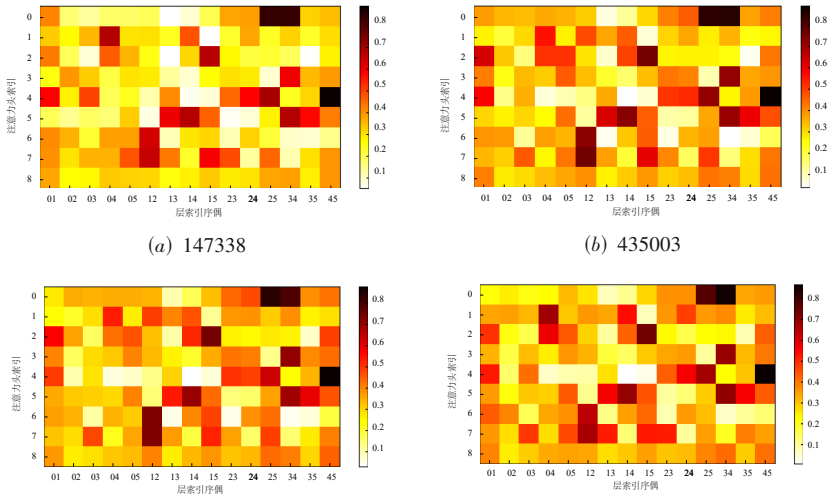
图A1　解码器各层各交叉注意力头权重与各层所有交叉注意力头权重均值的相似性均值

为进一步从局部视角定量描述误差传递，本文计算 Conditional DETR 模型解码器不同层各交叉注意力头的权重相似性，计算过程如式(A2)所示，<m,n>表示解码器层索引序偶，满足$m<n$，$m\in\{0,1,2,3,4\}$，$n\in\{1,2,3,4,5\}$，q为注意力头索引，解码器每层包含 8 个头，相似性度量函数 similarity(·)与式(A1)同，$\boldsymbol{w}_{m,q}$和$\boldsymbol{w}_{n,q}$分别为第m+1 层的第q+1 头和第n+1 层的第q+1 头的注意力权重，式(A2)的$\Omega_{<m,n>}$保留相似性结果.

$$\Omega_{<m,n>}=\{\text{similarity}(\boldsymbol{w}_{m,q},\boldsymbol{w}_{n,q})\}_{q\in\{0,1,\cdots,7\}} \tag{A2}$$

式(A2)计算结果的可视化如图 A2(a)~(d)，横坐标为 Conditional DETR 解码器层索引序偶，如 25 表示第 3 层和第 6 层，纵坐标 0~7 表示注意力头索引，纵坐标 8 索引行表示相应序偶层对应的所有注意力头权重相似性均值，每幅图右侧为图例，颜色越深表示相似性越强.

在图 A2 中，除索引行 8 外，每行表示不同序偶层同一位置注意力头的权重相似性，每列表示相应序偶层中各对同一位置的不同注意力头之间的权重相似性. 观察图 A2 可知，当不同样本输入 Conditional DETR 推理时，存在部分位置注意力头权重的层间相似性基本保持不变. 例如，解码器第 3 层与第 6 层的第 1 个注意力头的权重相似性，即(25,0)位置，解码器第 5 层与第 6 层的第 5 个注意力头的权重相似性，即(45,4)位置等. 我们认为这些基本不变的相似性关系是在辅助模型提取样本不变表征. 当然，大部分注意力头权重的层间相似性随输入样本不同而变化，说明各层注意力头权重与样本间可判别特征提取有关. 由图 A2 中各图索引行 8 的分布可知，在不同输入样本情况下，解码器中序偶层 25、34 和 45 之间所有注意力头权重相似性均值明显高于其他序偶层，与图 3 和图 A1 结果一致，说明各注意力头权重影响模型性能. 在图 A2 中，序偶区 1~5 反映出 Conditional DETR 解码器第 1 层与其余层对应位置的注意力头权重存在联系，图 3 已定性解释了解码器第 1 层输出特征存在明显偏差，而解码器特征提取与各注意力头权重有关，因此，第 1 层各注意力头权重中的偏差可层层传递.



(a) 147338 (b) 435003

(c) 571598 (d) 581317

图 A2 解码器不同层各注意力头权重相似性

附录 B 各组件消融实验

本文解耦基线模型中交叉注意力计算过程，分别提出 IAAWR 方法和值向量 AFE 方法提升注意力表现，同时提出 IRPR 方法细化预测框. 本节展示各方法的消融实验.

B1 1IAAWR 方法消融实验

IAAWR 方法包含表 B1 所示的 6 种方式，其中基于加权软机制 4 种，无加权硬机制 2 种. 方式 I 如式(B1)所示，其中①~⑤位置索引与表 B1 中①~⑤索引列对应，方式 II~IV 表达式结构与式(B1)相同，仅需将各参数填充在对应位置. 方法 V 和 VI 表示②和④位置直接相加得到⑤位置值.

$$\underset{⑤}{\boldsymbol{w}'_{j,k}}=\underset{①}{\alpha_j}\times\underset{②}{\boldsymbol{w}_{j,k}}+\underset{③}{(1-\alpha_j)}\times\underset{④}{\boldsymbol{w}'_{j-1}} \tag{B1}$$

在表 B1 中，j和k分别为层索引和注意力头索引，其中$j\in\{1,2,3,4,5\}$，$k\in\{0,1,2,3,4,5,6,7\}$；α_j为第j+1 层可学习权重，由于各层权重正相关，故$\alpha_j\in(0,1)$，$\boldsymbol{w}_{j-1}$和$\boldsymbol{w}_j$分别为第j层和第j+1 层权重均值；$\boldsymbol{w}'_j$为第j+1 层更新后的权重均值；$\alpha_{j,k}$为第j+1 层第k+1 注意力头的可学习权重，$\alpha_{j,k}\in(0,1)$；$\boldsymbol{w}_{j-1,k}$和$\boldsymbol{w}_{j,k}$分别为第j层第k+1 头和第j+1 层第k+1 头的注意力权重；$\boldsymbol{w}'_{j,k}$为更新后第j+1 层中

表B1　层间交叉注意力权重迭代更新的不同方法

机制	方法	①	②	③	④	⑤
软机制	Ⅰ	α_j	$\boldsymbol{w}_{j,k}$	$(1-\alpha_j)$	$\boldsymbol{w}'_{j-1}$	$\boldsymbol{w}'_{j,k}$
	Ⅱ	α_j	$\boldsymbol{w}_{j,k}$	$(1-\alpha_j)$	$\boldsymbol{w}'_{j-1,k}$	$\boldsymbol{w}'_{j,k}$
	Ⅲ	$\alpha_{j,k}$	$\boldsymbol{w}_{j,k}$	$(1-\alpha_{j,k})$	$\boldsymbol{w}'_{j-1,k}$	$\boldsymbol{w}'_{j,k}$
	Ⅳ	$\alpha_{j,k}$	$\boldsymbol{w}_{j,k}$	$(1-\alpha_{j,k})$	$\boldsymbol{w}'_{j-1}$	$\boldsymbol{w}'_{j,k}$
硬机制	Ⅴ	—	$\boldsymbol{w}_{j,k}$	—	$\boldsymbol{w}'_{j-1}$	$\boldsymbol{w}'_{j,k}$
	Ⅵ	—	$\boldsymbol{w}_{j,k}$	—	$\boldsymbol{w}'_{j-1,k}$	$\boldsymbol{w}'_{j,k}$

第k+1头的注意力权重.

使用表B1中IAAWR方法的不同构成方式改进基线模型中解码器各层交叉注意力权重计算过程，使用MS-COCO训练集微调训练模型，MS-COCO验证集测试模型，测试结果如表B2所示，各评价指标如正文4.1节所述.

由表B2可知，方式Ⅰ在除APS指标外的其他指标上得到较好的结果. 相对基线模型，该方式仅引入了5个可学习参数. 方式Ⅰ和Ⅱ的实验结果对比表明：每层各注意力头在相同权重系数时，若各注意力头按方法Ⅱ独立学习，会导致模型收敛缓慢. 方式Ⅲ和Ⅳ性能相当，说明不同注意力头使用不同权重系数时，对更新权重过程中所用的基底不敏感，但均不及引入参数量更少的方式Ⅰ. 实验结果对比可见，硬机制方式Ⅴ和Ⅵ的整体性能均不及对应的软机制方式Ⅰ和Ⅲ，说明了加权软机制对IAAWR方法的必要性. 因此，本文采用IAAWR方法的方式Ⅰ改进模型.

表B2　IAAWR方法不同变体改进的基线模型在MS-COCO验证集中的实验结果　　单位：%

机制	方式	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
软机制	Ⅰ	**40.8**	**61.0**	**43.5**	19.8	**44.5**	**59.3**
	Ⅱ	38.9	59.1	40.8	20.3	42.1	56.6
	Ⅲ	39.8	60.2	42.2	19.5	43.3	57.8
	Ⅳ	39.9	60.1	42.2	19.9	43.6	58.0
硬机制	Ⅴ	39.4	59.6	41.5	**20.8**	42.7	57.0
	Ⅵ	39.7	60.2	41.9	18.5	43.1	58.3

注：加粗字体表示最佳结果.

进一步验证不同的初始化参数对IAAWR方法的方式Ⅰ的影响. 使用Sigmoid函数约束IAAWR方法中各学习参数的取值范围在(0,1)，参数初始化计算方式如式(B2)所示，j为层索引，$j\in\{1,2,3,4,5\}$，α_j和init_value_j分别为第j+1层可学习权重(四舍五入保留两位小数)和初始化值，则IAAWR方法的方式Ⅰ改进的基线模型在MS-COCO验证集的实验结果如表B3所示，random指各参数随机初始化，fixed指各参数初始值固定为0.

$$\alpha_j=\text{Sigmoid}(\text{init\_value}_j) \tag{B2}$$

表B3　IAAWR方法的方式Ⅰ的不同初始化参数对应基线模型在MS-COCO验证集的实验结果　　单位：%

init_value_j	α_j	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
−2	≈0.12	37.1	57.0	38.8	14.2	40.2	58.0
−1	≈0.27	36.8	56.7	38.5	14.6	40.2	57.6
0	0.50	**40.8**	**61.0**	**43.5**	**19.8**	**44.5**	**59.3**
1	≈0.73	37.2	57.1	39.1	15.4	40.4	58.2
2	≈0.88	38.0	58.2	39.9	15.3	41.1	59.5
random	random	36.9	56.8	38.8	14.8	40.3	57.4
fixed	fixed	39.3	59.3	41.6	19.1	42.9	56.4

注：加粗字体表示最佳结果.

表B3结果表明：IAAWR方法对参数初始化值敏感，各层参数初始值不同，模型效果不同. 各参数随机初始化的效果劣于固定初始化的效果，说明随机初始化不利于模型收敛于局部最佳. 当各层参数初始值为0时，随训练过程动态学习优于保持固定不变，说明可学习对IAAWR方法的重要性. 当选取−2、−1、0、1和2作为初始值时，各层初始值为0，即权重为0.5时，模型达到最佳效果. 因此，本文改进模型即约束IAAWR方法

各层参数可学习，且初始值为0.

B2 值向量AFE方法消融实验

如正文式(7)所示，值向量AFE方法引入了6个可学习参数，动态增强了基线模型编码器各层的输出特征，表B4为值向量AFE方法不同参数初始化情况下改进的基线模型，在MS-COCO验证集的实验结果，random表示各参数随机初始化，fixed为各参数初始值为0，训练过程中保持固定，β_t对应编码器第t+1层输出特征加权值，S表示S型函数，这里指Sigmoid或Tanh函数，表示init_value$_t$到β_t的映射，即

$$\beta_t = S\left(\text{init\_value}_t\right) \tag{B3}$$

其中，$t \in \{0, 1, 2, 3, 4, 5\}$. 在表B4中，第1行说明：当模型编码器各层输出特征直接相加作为增强特征时，效果弱于加权增强的方法. 当S型函数采用Sigmoid或Tanh时，改进模型效果没有明显差异，说明值向量AFE方法对参数初始化方式和S型函数选择不敏感. 当S型函数使用Sigmoid时，可看出值向量AFE方法随机参数初始化的效果弱于固定参数，说明随机参数初始化不易收敛至局部最佳；固定参数弱于参数可学习的情况，说明可学习权重是值向量AFE方法提升模型性能的关键. 鉴于Sigmoid搜索空间更小，因此本文采用Sigmoid函数，初始值为0.00，β_t为0.5的设置.

表B4 值向量AFE参数初始化在MS-COCO验证集的消融实验结果 单位：%

S	init_value$_t$	β_t	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
—	+∞	1.00	39.0	59.7	41.1	18.5	42.5	57.0
Sigmoid	0.00	0.50	**41.1**	**61.4**	**43.4**	**21.0**	**44.9**	58.9
	random	random	40.6	60.9	42.9	20.0	44.3	58.9
	fixed	fixed	40.9	61.2	43.3	21.1	44.4	58.5
Tanh	0.50	≈0.50	41.0	61.3	**43.4**	20.0	44.5	**59.2**
	random	random	40.8	61.2	43.3	20.2	44.7	**59.2**

注：加粗字体表示最佳结果.

值向量AFE方法通过对基线模型编码器各层输出特征加权，增强交叉注意力计算过程中的值向量语义，由于基线模型中编码器-解码器结构相似，值向量AFE方法也可用于对解码器输出特征的增强. 表B5展示了当值向量AFE方法应用于基线模型不同位置时，在MS-COCO验证集中的实验结果. 其中，第1列表示AFE添加的位置，“—”表示基线模型，decoder和encoder分别表示AFE分别用于增强解码器和编码器的各层输出特征，encoder&decoder表示值向量AFE方法同时用于增强解码器和编码器各层输出特征，decoder improvement表示当值向量AFE方法用于解码器输出特征增强时，不考虑解码器第1层的输出特征.

表B5 基线模型不同位置嵌入值向量AFE方法在MS-COCO验证集实验 单位：%

添加位置	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
—	40.1	60.2	42.4	19.6	43.5	58.7
decoder	39.6	59.9	41.8	19.9	42.7	57.3
decoder improvement	40.7	61.2	43.0	20.3	44.3	58.9
encoder	41.1	61.4	43.4	**21.0**	**44.9**	58.9
encoder&decoder	**41.3**	**61.6**	**43.9**	20.5	44.7	**60.1**

注：加粗字体表示最佳结果.

表B5实验结果表明：值向量AFE方法同时增强基线模型编码器和解码器输出特征时(encoder&decoder)，评价指标AP较基线模型提升最大，值向量AFE方法用于基线模型编码器时(encoder)，评价指标AP提升程度次之，明显优于值向量AFE方法用于增强解码器输出特征时(decoder)的模型性能. 在值向量AFE方法引入encoder&decoder的情况下，引入可学习参数量是值向量AFE方法仅引入encoder情况下的2倍，对基线模型的性能增益与仅引入到encoder的情况接近，因此，本文仅将AFE用于增强基线模型的encoder. 值向量AFE方法用于解码器特征增强时不考虑解码器第1层输出特征的情况(decoder improvement)优于值向量AFE方法用于增强解码器各层特征(decoder)的表现，进一步证实了本文阐述的模型解码器第1层存在学习偏差，易误导模型得到次优结果.

B3 IRPR方法消融实验

文中式(8)和式(9)表示在IRPR方法计算过程，当layer$_u$在模型解码器各层参考点计算过程中参数不共享时，由于解码器各层对应的参考点生成网络参数随机初始化，生成的参考点随机性大，导致模型训练崩溃. 因此，消融实验中仅考虑共享layer$_u$的情况. 在MS-COCO验证集的消融实验结果如表B6所示. Detach表示当前层参考

点反向传播时是否考虑上一层参考点梯度的计算，×和√分别表示考虑和不考虑上一层参考点对梯度的影响.

表B6　IRPR方法在MS-COCO验证集的消融实验结果　　单位：%

Detach	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
×	39.2	58.8	41.6	19.7	42.5	57.4
√	**40.4**	**60.5**	**43.1**	**19.8**	**43.6**	**58.3**

注：加粗字体表示最佳结果.

表B6消融实验结果表明：当前层参考点产生的梯度更新网络参数时，不考虑前一层参考点对梯度的影响，效果更优，若当前层和上一层参考点同时计算梯度，更新网络参数，容易导致训练不稳定，延缓模型收敛速度.

附录C　以方式I为例证明IAAWR方法的有效性

本节详细展示正文式(5)和式(6)的推导过程，并结合推导过程，证明文中提出的IAAWR方法中方式I的有效性，IAAWR方法其他方法可依此类比证明. 正文式(5)和式(6)推导过程如下：

当$J=0$时，由式(4)可得$\boldsymbol{w}_0'=\boldsymbol{w}_0$. 当$J=1$时，由式(4)可得：

$$\boldsymbol{w}_1'=\alpha_1\times\boldsymbol{w}_1+(1-\alpha_1)\times\boldsymbol{w}_0 \tag{C1}$$

当$J=2$时，由式(4)可得：

$$\boldsymbol{w}_2'=\alpha_2\times\boldsymbol{w}_2+(1-\alpha_2)\times\boldsymbol{w}_1' \tag{C2}$$

将式(C1)带入式(C2)，可得：

$$\begin{aligned}\boldsymbol{w}_2'&=\alpha_2\times\boldsymbol{w}_2+(1-\alpha_2)\times[\alpha_1\times\boldsymbol{w}_1+(1-\alpha_1)\times\boldsymbol{w}_0]\\&=\alpha_2\times\boldsymbol{w}_2+\alpha_1\times\boldsymbol{w}_1+\boldsymbol{w}_0-\alpha_1\times\boldsymbol{w}_0-\alpha_2\times\alpha_1\times\boldsymbol{w}_1\\&\quad-\alpha_2\times\boldsymbol{w}_0+\alpha_2\times\alpha_1\times\boldsymbol{w}_0\\&=\alpha_2\times(\boldsymbol{w}_2-\boldsymbol{w}_0)+\alpha_1\times(1-\alpha_2)\times(\boldsymbol{w}_1-\boldsymbol{w}_0)+\boldsymbol{w}_0\end{aligned} \tag{C3}$$

当$J=3$时，由式(4)可得式(C4)：

$$\begin{aligned}\boldsymbol{w}_3'&=\alpha_3\times(\boldsymbol{w}_3-\boldsymbol{w}_0)+\alpha_2\times(1-\alpha_3)\times(\boldsymbol{w}_2-\boldsymbol{w}_0)\\&\quad+\alpha_1\times(1-\alpha_2)\times(1-\alpha_3)\times(\boldsymbol{w}_1-\boldsymbol{w}_0)+\boldsymbol{w}_0\end{aligned} \tag{C4}$$

以此类推，任意层更新后的交叉注意力权重均值如式(C5)和(C6)所示，即正文中的式(5)和式(6)，层索引$J\in\{1,2,3,4,5\}$.

$$\boldsymbol{w}_J'=\sum_{j=1}^{J}\lambda_j\times(\boldsymbol{w}_j-\boldsymbol{w}_0)+\boldsymbol{w}_0 \tag{C5}$$

$$\lambda_j=\begin{cases}\alpha_j\times(1-\alpha_{j+1})\times\cdots\times(1-\alpha_J), & 1\leqslant j<J\\ \alpha_j, & j=J\end{cases} \tag{C6}$$

由于$\alpha_j\in(0,1)$，结合式(C6)，可得$\lambda_j\in(0,1)$，进一步可证$\sum_{j=1}^{J}\lambda_j\in(0,1)$，具体证明如下：

当J=1时，由式(C6)可得$\lambda_1=\alpha_1$，因此$\lambda_1\in(0,1)$. 当J=2时，由式(C3)、(C5)和(C6)可知：

$$\begin{aligned}&\sum_{j=1}^{2}\lambda_j=\alpha_2+\alpha_1\times(1-\alpha_2)\\&\because\quad \alpha_j\in(0,1),\qquad \therefore\ \sum_{j=1}^{2}\lambda_j\in(0,1).\end{aligned} \tag{C7}$$

当J=3时，由式(C4)、式(C5)和式(C6)可知：

$$\begin{aligned}\sum_{j=1}^{3}\lambda_j&=\alpha_3+\alpha_2\times(1-\alpha_3)\\&\quad+\alpha_1\times(1-\alpha_2)\times(1-\alpha_3)\\&=\alpha_3+[\alpha_2+\alpha_1\times(1-\alpha_2)]\times(1-\alpha_3)\\&\because\quad \alpha_j\in(0,1),\qquad \therefore\ \sum_{j=1}^{3}\lambda_j\in(0,1).\end{aligned} \tag{C8}$$

以此类推，证得$\sum_{j=1}^{J}\lambda_j\in(0,1)$. 因此，在式(C5)中，当$\sum_{j=1}^{J}\lambda_j$趋于0时，$\boldsymbol{w}_J'$趋近$\boldsymbol{w}_0$；当$\sum_{j=1}^{J}\lambda_j$趋于1时，$\boldsymbol{w}_J'$可避免$\boldsymbol{w}_0$的影响；当$\sum_{j=1}^{J}\lambda_j$在0~1时，随着$\sum_{j=1}^{J}\lambda_j$的不断增大，可逐步削弱$\boldsymbol{w}_0$对$\boldsymbol{w}_J'$的影响.

根据此性质，可得结论：网络可动态学习解码器第1层与其余各层的关系.

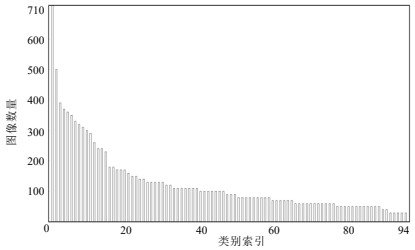
结合正文图7和上述推导结论可知，本文提出的IAAWR方法是有效的.

附录D　Butterfly_2018和Butterfly_2023

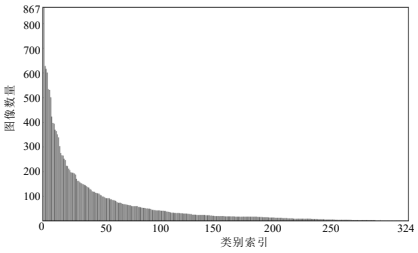
Butterfly_2018源于“2018年第三届中国数据挖掘竞赛——国际首次蝴蝶识别大赛”的竞赛数据集，包含94类蝴蝶，生态照1 408张和标本照486张，各类生态照数量按照1∶1划分成训练集和测试集，训练集数量向上取整，标本照全部置于训练集，最终训练集包含1 207(721+486)张图像，测试集包含687张蝴蝶生态照，该划分过程与文献[16]一致. 同时，为保证深度学习的训练数据需求，训练集对每张图片进行离线随机扩增(随机翻转、随机裁剪、随机加噪等)9倍，扩增后训练集图像数量为12 070张，测试集不进行数据扩增. Butterfly_2023为实验室野外采集的私有蝴蝶生态照数据集，含324类蝴蝶共21 728张图像，为保证训练集图像数量，若某类蝴蝶图像数量低于10张，则随机挑选1张置于测试集，否则，随机挑选该类图像总数的1/10(下取整)置于测试集，其余归入训练集，最终训练集包含19 601张图像，测试集包含2 127张图像.

Butterfly_2018和Butterfly_2023训练集中各类图像数量的分布情况分别如图D1所示，横坐标为类索引，纵坐标为图像数量.

图D1直观反映了实验用的2个蝴蝶数据集的各类



(*a*) Butterfly_2018 训练集



(*b*) Butterfly_2023 训练集

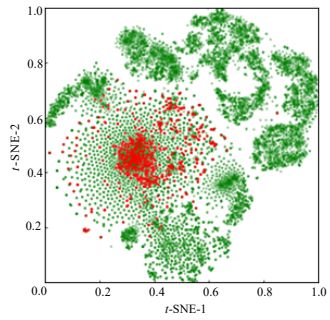
图 D1 Butterfly_2018 和 Butterfly_2023 训练集各类蝴蝶图像数量分布

蝴蝶图像数量呈现明显长尾分布，由于这 2 个数据集中各图像包含的目标实例数量基本相同，各类别实例数量分布情况与图 D1(*a*)~(*b*)类似，故 Butterfly2018 和 Butterfly_2023 为类间样本量不平衡数据集. 这 2 个蝴蝶数据集的训练集中各类图像数量最大值与最小值之比分别为 23.7 和 867.0，Butterfly_2023 的不平衡性更厉害，是更具有挑战性的数据集.

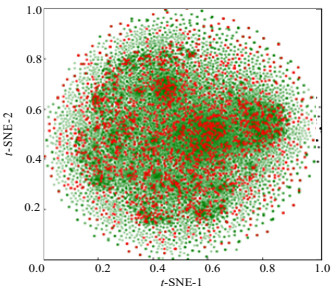
为了进一步了解数据集，我们对蝴蝶图像数据集 Butterfly_2018 和 Butterfly_2023 中各图像采用 *t*-SNE (*t*-distributed Stochastic Neighbor Embedding)降维，图 D2(*a*)和图 D2(*b*)为这 2 个数据集降维后结果的可视化，绿色点表示训练子集样本，红色点为测试子集样本.

由图 D2 可知，Butterfly_2018 数据集的训练集和测试集样本分布较 Butterfly_2023 有更大差异，这是由于 Butterfly_2018 的训练集包括野外环境蝴蝶图像和标本蝴蝶图像两类，且采用了多种图像增强技术，而测试集仅包含野外环境下的蝴蝶图像. Butterfly_2023 的训练集和测试集均为野外环境下的蝴蝶图像.

综合图 D1~图 D2，Butterfly_2018 和 Butterfly_2023



(*a*) Butterfly_2018 数据集



(*b*) Butterfly_2023 数据集

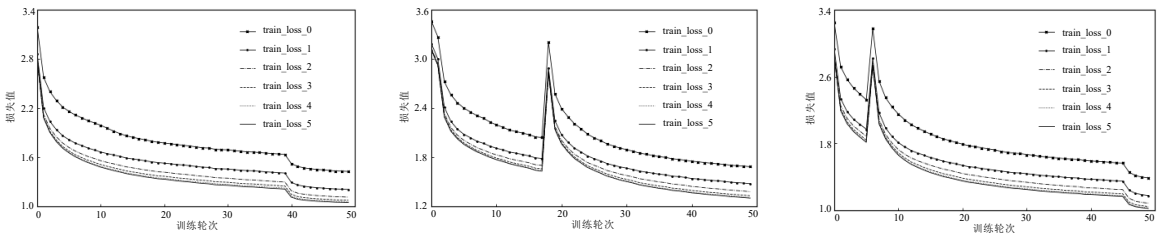
图 D2 Butterfly_2018 和 Butterfly_2023 用 *t*-SNE 降维后的可视化结果

较 MS-COCO 的样本分布更复杂，识别难度更大，是更具挑战性的数据集.

附录 E Conditional DETR 源码中不同 backbone 的损失函数日志

本节将 Conditional DETR 源码中 backbone 为 DC5-ResNet50、ResNet101 和 DC5-ResNet101 的三种不同 Conditional DETR 的解码器训练损失函数可视化，如图 E1 所示，横坐标为训练轮次，纵坐标为损失值.

结合图 E1 和图 3，说明模型训练损失函数的变化与模型所选 Backbone 有关，当 Backbone 为 ResNet50 时，模型训练损失呈整体下降趋势，当 Backbone 为 ResNet101 时，模型训练损失的趋势为当下降至某轮次时出现阶跃，最后再下降. 此外，ResNet 最后 1 个阶段采用空洞卷积(Dilated Convolution，DC5)可辅助模型收敛至更小的损失值.



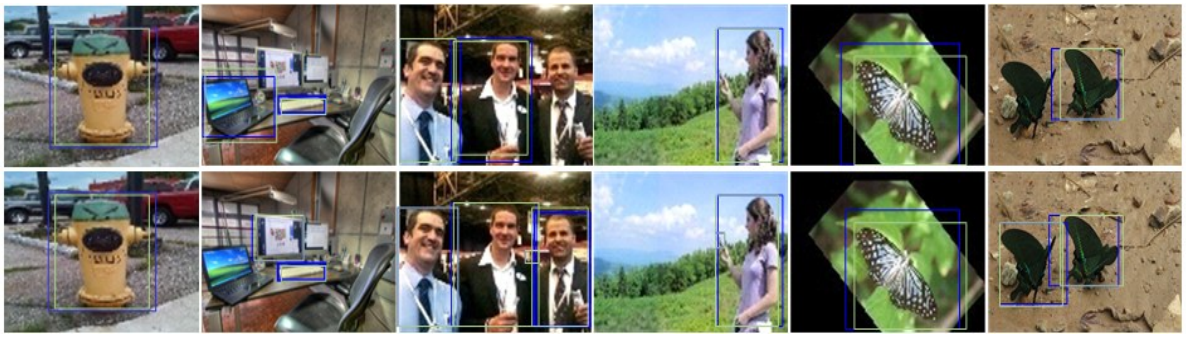
(*a*) Backbone 为 DC5-ResNet50　(*b*) Backbone 为 ResNet101　(*c*) Backbone 为 DC5-ResNet101

图 E1 Backbone 为 DC5-ResNet50、ResNet101 和 DC5-ResNet101 的 Conditional DETR 的解码器不同层的训练损失变化曲线

另外，由图E1和图3的训练损失曲线揭示，不同Backbone的Conditional DETR的解码器第1层输出的训练损失值（train_loss_0曲线）远高于其余层，说明第1层输出与期望输出之间存在较大偏差．解码器层层堆叠，各层输出偏差层层传递，逐层递减，第5、6层（train_loss_4和train_loss_5曲线）的输出偏差几乎重合，且是各层输出的偏差中最小的．再次验证了本文研究的意义和正确性，以及提出IAAWR方法来动态调节解码器不同层的交叉注意力权重，削弱学习偏差层间传递的意义和必要性．

附录F　检测结果

本节将展示模型的检测效果．图F1展示了不同模型分类置信度高于0.9的检测结果，第1行对应基线模型检测结果，第2行对应本文改进模型检测结果，第1~4列为MS-COCO验证集图像，第5~6列分别为Butterfly_2018和Butterfly_2023验证集图像，图中蓝色为预测框，绿色为真实框，模型未预测出的目标，不标注真实框．



图F1　基线模型Conditional DETR和本文AI-DETR模型在MS-COCO、Butterfly_2018和Butterfly_2023数据集的检测结果

图F1的检测结果可知，本文AI-DETR模型在稠密或遮挡情况下的检测能力优于Conditional DETR基线模型，如图F1第3列中的“人”目标的检测和第4列中“手机”的检测，以及第6列中的左侧“蝴蝶”目标检测，AI-DETR检测出了这3张图像中基线模型没有发现的人、手机和蝴蝶．此外，在目标边界不清晰时，如图F1第6列，本文提出的AI-DETR模型能够有效区分边界，发现多个目标．在预测框的精细度方面，本文AI-DETR模型也表现出明显优势，如图F1第5列中“蝴蝶”的检测，本文AI-DETR模型预测出的目标边框更接近真实框．然而，在某些实例的检测方面，尤其是当图像包含多个实例时，对于相对较大实例的检测，本文AI-DETR模型劣于基线模型．例如，图F1第2列中的黑色电脑屏，但AI-DETR模型发现了另一个基线模型没有发现的较小目标（远处的电脑显示屏）．这也验证了本文AI-DETR模型在目标召回率、小目标识别和预测框的精细化方面优于基线模型．

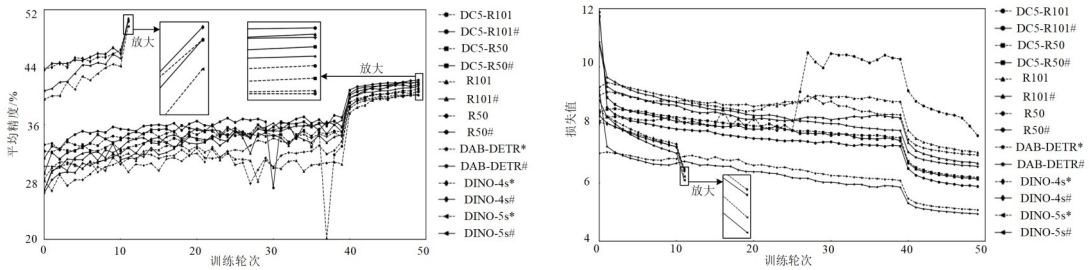
附录G　AI-DETR模型收敛性、检测性能和通用性测试

本节的G1节将测试正文4.3节各改进模型的收敛性，验证本文创新点对模型收敛性的影响．G.2小节将本文AI-DETR模型与近期DETR改进方法进行对比，验证本文少参数优化DETR方案的优越性．G.3小节将在Butterfly_2018和Butterfly_2023数据集中比较本文提出的AI-DETR模型与多种通用目标检测方法，以测试AI-DETR模型在自然环境中对蝴蝶目标检测的优势．G.4小节引入4种不同领域数据集，测试本文AI-DETR模型对不同领域目标检测任务的普适性．

G1　本文方法改进模型的收敛性测试

为验证提出的3个创新点对模型收敛性的影响，图G1（a）和图G1（b）分别展示了4.3节提及的本文创新点改进的各模型与对应基线模型在MS-COCO验证集的AP值（纵坐标）随训练轮次（横坐标）的变化曲线，以及在MS-COCO训练集的总损失变化曲线（横坐标为训练轮次，纵坐标为损失值）．图例对应正文表3~表5中各模型．

由图G1（a）和图G1（b）可知，各模型随训练轮次增加，验证集平均精度逐步增加，训练集损失逐步降低，说明各模型训练过程有效．其中，以DC5-R101为backbone的Conditional DETR模型在训练过程中震荡最为明显，本文改进模型DC5-R101#有效提升了基线模型训练稳定性．其余改进模型与基线模型训练变化曲线趋势相仿．除训练初期，图G1（a）中本文方法改进模型高于对应的基线模型，即改进模型在验证集的平均精度高于对应的基线模型，图G1（b）中本文方法改进模型低于对应的基线模型，即改进模型在训练集损失低于对应的基线模型．综合图G1（a）和图G1（b）各模型的AP和Loss曲线的变化趋势，说明本文方法可提升模型训练收敛性．图G1（a）和图G1（b）各模型的AP和Loss变化曲线震荡性整体弱于对应的基线模型，进一步说明本文方法改进的模型在检测性能和收敛性方面均优



(*a*) 各模型在MS-COCO验证集的平均精度变化曲线 (*b*) 各模型在MS-COCO训练集的损失变化曲线

图G1 各改进模型与对应基线模型的平均精度和训练损失变化曲线

于对应的基线模型,说明提出方法对模型的性能带来积极影响.

G2 本文AI-DETR模型与其他DETR模型的性能比较

为验证本文提出方法的先进性,表G1展示本文创新点改进后模型与其他DETR模型(单尺度和多尺度)在MS-COCO验证集的检测结果. 表中各模型骨干网络均为ResNet50,这里为了描述方便,将本文创新点改进后的Conditional DETR记为AI-DETR-S(即正文的AI-DETR),*S*表示使用骨干网络的单尺度特征,将本文方法改进后的DINO-4s记为AI-DETR-M(即正文表4的DINO-4s#),*M*表示使用骨干网络的多尺度特征. 为与本文方法公平比较,表G1选择的单尺度基线检测模型主要涉及DETR解码器中注意力机制的改进. 表G1多尺度检测模型主要使用骨干网络输出的多尺度特征图作为DETR编码器的输入. 含"√"的行表示多尺度检测方法.

表G1 本文AI-DETR模型与其他DETR改进方法在MS-COCO验证集的实验结果

模型名称	训练轮次	多尺度检测	AP/%	AP_{50}/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%	GFLOPs/G	FPS/s
Deformable DETR[7]	50	—	39.4	59.6	42.3	20.6	43.0	55.5	78	7
Anchor DETR[31]	50	—	38.5	61.5	39.8	16.7	43.1	59.2	160	8
SMCA[32]	50	—	39.2	60.8	41.9	20.3	43.0	55.2	86	10
Conditional DETR[8]	50	—	40.1	60.2	42.4	19.6	43.5	58.7	90	14
DAB-DETR[9]	50	—	40.1	60.2	42.4	17.5	43.8	61.5	94	14
AI-DETR-S(Ours)	40	—	40.3	60.5	42.9	20.3	43.4	58.3	90	13
AI-DETR-S(Ours)	50	—	41.9	62.2	44.3	21.5	45.1	60.4	90	13
Deformable DETR[7]	50	√	43.8	62.6	47.7	26.4	47.1	58.0	173	6
SMCA[32]	50	√	43.7	63.6	47.2	24.2	47.0	60.4	152	8
TSP-RCNN[33]	36	√	40.0	59.4	43.8	24.5	43.2	51.5	188	7
TSP-FCOS[33]	36	√	42.8	62.0	46.6	25.9	46.4	55.5	189	9
RT DETR[34]	12	√	48.0	65.6	52.1	29.0	52.9	**65.5**	136	25
DINO-5s[10]	12	√	49.7	66.7	54.2	33.4	52.5	64.0	860	9
DINO-4s[10]	12	√	50.4	67.9	55.1	32.5	**54.0**	64.3	279	16
AI-DETR-M(Ours)	12	√	**50.7**	**68.4**	**55.2**	**33.8**	53.7	65.1	279	15

注:加粗字体表示最佳结果.

表G1展示的各单尺度检测方法在MS-COCO验证集的实验结果表明:Deformable DETR采用局部关键点计算注意力,对小尺寸目标检测精度较高,Anchor DETR采用行列解耦注意力机制,对大尺寸目标检测效果显著,空间调制协同注意力(Spatially Modulated Co-Attention,SMCA)采用高斯分布的空间调制协同注意力,倾向提升DETR小尺寸目标检测能力. Conditional DETR和DAB-DETR将空间和语义解耦,改进DETR中的交叉注意力机制,检测精度和推理速度(FPS)远高于Deformable DETR、Anchor DETR和SMCA,计算量(GFLOPs)适中. 其中,Conditional DETR对不同尺寸目标的检测能力更均衡. 本文AI-DETR-S基于Conditional DETR,仅引入11个额外加权参数,对计算量影响可忽略不计,推理速度与基线模型相当,与其他单尺度检测模型训练轮次相同(50轮)时,AI-DETR-S检测结果最佳. 经40轮训练,AI-DETR-S就可比其他单尺度检测模型性能更好,这也再次验证了本文方法对模型训练收敛性的积极影响.

表G1各多尺度检测方法的检测结果揭示，Deformable DETR和SMCA均改进DETR解码器中的交叉注意力机制，较单尺度模型AI-DETR-S在评价指标AP_{75}、AP_S和AP_M中提升明显，但计算量显著增大，推理速度降低. 基于Transformer的集合预测与RCNN(Transformer-based Set Prediction with RCNN，TSP-RCNN)和基于Transformer的集合预测与全卷积单阶段目标检测(Transformer-based Set Prediction with Fully Convolutional One-Stage object detection，TSP-FCOS)在骨干网络与DETR编码器之间引入空间引导的特征筛选机制，将潜在目标特征作为DETR编码器输入，编码器输出与预测头相连，实现训练仅36轮次便快速收敛，但模型精度不高. 实时检测Transformer(Real-Time Detection Transformer，RT-DETR)采用混合编码器和基于交并比(Intersection over Union，IoU)的查询集感知机制，虽推理速度快，计算量小. 然而，同样训练12轮次，精度低于本文模型A-DETR-M. DINO-4s和DINO-5s分别将骨干网络输出的4个和5个尺寸的特征图作为输入，采用多种策略联合训练，在MS-COCO验证集精度分别为50.4%和49.7%. AI-DETR-M将本文提出的值向量AFE方法用于DINO-4s，同样经过12轮次训练，在保证推理速度前提下，精度高达50.7%，小目标检测精度较DINO-4s提升1.3个百分点，进一步验证本文方法对模型检测小目标的能力提升明显.

表G1的实验结果明显可知，本文提出的模型AI-DETR-S和AI-DETR-M分别在各比较的单尺度和多尺度检测方法中，检测精度达到最佳. 单尺度检测方法的比较结果还表明：本文方法可提升模型收敛速度. 各多尺度检测方法的实验结果进一步验证了本文方法提升了改进模型AI-DETR-M的小目标检测效能.

G3　本文AI-DETR模型在蝴蝶数据集Butterfly_2018和Butterfly_2023的实验测试

表G2展示了更多目标检方法在Butterfly_2018和Butterfly_2023数据集的实验结果. 文献[21，23，24]是基于Butterfly_2018数据集提出的蝴蝶检测领域的先进Anchor-Based检测方法. 本文提出的AI-DETR本质上是单阶段Anchor-Free的目标检测方法，因此表G2中其他对比方法均为单阶段Anchor-Free的目标检测方法. 各检测方法均遵从默认配置. 其中，KSRFB-net_50和MRFB-net_50分别表示文献[23，24]提出的模型在Butterfly_2018训练集训练50轮得到的最佳模型. AI-DETR-S是本文提出的3个创新点改进的Conditional DETR，即正文中的AI-DETR. AI-DETR-M是本文创新点改进后的DINO-4s，即正文表4的DINO-4s#.

在表G2中，相同模型在Butterfly_2018数据集的检测性能弱于在Butterfly_2023数据集的检测性能，主要是由于Butterfly_2018数据集源域和目标域分布偏移更大. DETR系列方法的推理速度(FPS)显著高于其他方法，主要由于其他方法采用了处理相对耗时的锚点密集采样策略和非极大值抑制(Non-Maximum Suppression，NMS)后处理技术.

在Butterfly_2018数据集中，多尺度目标检测方法(RepPoints、SparseRCNN和Deformable DETR等)检测效果显著高于单尺度目标检测方法(Conditional DETR和DAB-DETR)，说明多尺度特征对跨域目标检测问题有一定帮助. 本文提出的单尺度目标检测方法AI-DETR-S在Butterfly_2018数据集高于基线模型(Conditional DETR)1.3个百分点，性能接近多尺度检测方法(FoveaBox). 多尺度检测方法AI-DETR-M在AP和AP_{75}两个指标达到最佳，较基线模型(如DINO)分别高出1.3个百分点和0.3个百分点. KSRFB-net与MRFB-net[23,24]在AP_{50}指标具有显著优势，除与使用Anchor-Based的模型及并行特征提取结构有关外，可能还与训练轮次较多有关(分别为200和280)，在表G2中，KSRFB-net_50和MRFB-net_50分别表示KSRFB-net与MRFB-net[23,24]在训练50轮时的模型. 它们的AP_{50}分别为80.4%和79.8%，低于AI-DETR-S(训练50轮)和AI-DETR-M(训练12轮)的AP_{50}性能值.

对Butterfly_2023数据集，本文提出的AI-DETR-M模型在所有评价指标达到最佳，AP指标较基线模型(如DINO)高出2.3个百分点. AI-DETR-S在各评价指标的值接近多尺度检测方法Deformable DETR的性能. 此外，基于生成的DiffusionDet在Butterfly_2023数据集取得次优的AP值81.7%，但在Butterfly_2018数据集中表现一般，说明该方法在同域图像检测方面存在优势，在跨域图像检测方面仍有提升空间. 基于卷积神经网络[35~41]的检测方法在Butterfly_2023数据集中的性能普遍低于DETR方法[7~10]，而在Butterfly_2018数据集中的性能与DETR方法接近，鉴于Butterfly_2018训练集中存在数据扩增带来的几何或色彩畸变，说明卷积的局部注意力在处理畸变图像时更具优势，DETR的全局注意力在处理规则图像时更优.

综上分析可见，本文提出的AI-DETR模型在富有挑战性的数据集Butterfly_2018和Butterfly_2013的检测性能较其他方法有显著优势.

G4　本文AI-DETR模型的通用性测试

本节测试本文提出的AI-DETR模型在多种通用数据集的检测性能，以测试其通用性. 通用数据集涵盖自动驾驶、遥感图像、医学影像和自然图像几大场景，表G3为4种通用数据集的详细信息.

KITTI(Karlsruhe Institute of Technology and Toyota technological Institute)、卫星影像多车辆数据集(Satellite Imagery Multivehicles Dataset，SIMD)和Brain-Tumor

表G2 本文提出的AI-DETR与现有新方法在Butterfly_2018和Butterfly_2023数据集的实验结果

数据集	模型名称	AP/%	AP_{50}/%	AP_{75}/%	FPS/s
Butterfly_2018	RetinaNet+DSEM[21]	65.0	81.2	72.3	4
	KSRFB-net[23]	—	88.1	—	3
	KSRFB-net_50[23]	—	80.4	—	3
	MRFB-net[24]	—	**89.1**	—	2
	MRFB-net_50[24]	—	79.8	—	2
	FSAF[35]	61.4	80.9	68.5	4
	ATSS[36]	61.9	76.6	67.2	4
	CenterNet[37]	64.8	81.1	71.3	4
	FoveaBox[38]	65.2	82.3	74.9	4
	AutoAssign[39]	63.7	80.6	71.3	4
	SparseRCNN[4]	66.7	80.9	73.0	5
	PAA[40]	66.2	81.3	73.6	4
	RepPoints[41]	67.3	84.3	76.4	4
	DiffusionDet[42]	65.8	82.0	74.1	1
	Conditional DETR[8]	63.7	82.1	72.2	6
	DAB-DETR[9]	63.5	80.7	71.2	6
	Deformable DETR[7]	66.7	83.7	74.9	5
	DINO[10]	70.7	84.5	76.7	6
	AI-DETR-S(Ours)	65.0	82.0	71.3	6
	AI-DETR-M(Ours)	**72.0**	85.5	**77.0**	6
Butterfly_2023	FSAF[35]	74.9	85.3	81.2	4
	ATSS[36]	70.3	80.5	76.6	4
	CenterNet[37]	70.9	81.5	78.6	4
	FoveaBox[38]	73.2	84.8	81.3	4
	AutoAssign[39]	68.3	80.3	76.5	4
	SparseRCNN[4]	78.6	87.9	84.8	5
	PAA[40]	71.1	80.6	77.3	4
	RepPoints[41]	70.2	81.3	77.3	4
	DiffusionDet[42]	81.7	91.5	88.2	1
	Conditional DETR[8]	78.9	89.2	85.7	6
	DAB-DETR[9]	79.1	90.1	86.5	6
	Deformable DETR[7]	80.5	90.2	86.7	5
	DINO[10]	81.2	90.1	86.6	6
	AI-DETR-S(Ours)	79.7	89.7	86.2	6
	AI-DETR-M(Ours)	**83.5**	**93.0**	**89.6**	6

注:加粗字体表示最佳结果.

为公开数据集,Butterfly_2025为实验室私有数据集.其中,KITTI数据集涉及真实场景下的不同类型车辆和不同姿态行人的细粒度检测.SIMD数据集为谷歌地球卫星拍摄的高分辨率遥感图像,涉及15种不同交通工具(车、船和飞机等)的细粒度检测,该数据集中单张图片中包含密集的实例分布.Brain-Tumor为脑肿瘤磁共振成像(Magnetic Resonance Imaging,MRI)横向切片数据集,涉及3种不同脑部肿瘤的检测.Butterfly_2025为实验室私有的蝴蝶生态图像数据集,除了将Butterfly_2018中的生态图像与Butterfly_2023所有数据合并外,融合新采集的254种蝴蝶生态图像,最终蝴蝶物种共包含631种.前3个数据集的训练集和测试集划分遵从发布者的默认划分,Butterfly_2025训练集和测试集划分方式与Butterfly_2023一致,见附录D.上述4种数据集均为分布不平衡的细粒度目标检测数据集,具有一定挑战性.

表G4展示了各模型对上述4种通用数据集的检测结果.鉴于实验资源限制,表G4除基线模型外,额外引

表G3　4种通用数据集详细信息

数据集名称	涉及领域	目标类别	训练集图像数量	测试集图像数量
KITTI[43]	自动驾驶	8	6 181	1 300
SIMD[44]	遥感影像	15	4 000	1 000
Brain-Tumor[45]	医学影像	3	6 930	990
Butterfly_2025	自然图像	631	36 665	3 983

表G4　各模型在4种通用数据集的实验结果　　单位：%

数据集名称	模型名称	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
KITTI[43]	SparseRCNN[4]	57.1	84.6	64.0	50.1	56.7	63.3
	Conditional DETR[8]	52.8	84.0	57.0	39.2	51.8	65.7
	DINO[10]	66.9	90.7	77.1	57.8	66.8	72.6
	AI-DETR-S	56.3	85.8	62.8	42.6	55.1	68.6
	AI-DETR-M	**67.6**	**91.7**	**77.3**	**57.9**	**67.4**	**73.8**
SIMD[44]	SparseRCNN[4]	58.4	75.8	68.3	7.9	54.0	63.2
	Conditional DETR[8]	66.2	83.1	77.7	9.8	59.0	73.9
	DINO[10]	69.1	85.6	80.6	11.9	61.5	**77.5**
	AI-DETR-S	68.0	84.1	78.4	11.2	60.8	74.6
	AI-DETR-M	**70.2**	**86.1**	**81.5**	**12.6**	**62.4**	76.5
Brain-Tumor[45]	Open Method[45]	—	79.7	—	—	—	—
	SparseRCNN[4]	37.9	70.6	35.8	15.4	53.2	57.0
	Conditional DETR[8]	47.8	78.8	50.5	22.5	66.8	88.1
	DINO[10]	46.6	76.8	48.5	21.1	63.4	84.4
	AI-DETR-S	**49.0**	**80.1**	**51.6**	**23.0**	**68.7**	**89.2**
	AI-DETR-M	46.8	77.0	48.8	21.4	63.4	84.4
Butterfly_2025	SparseRCNN[4]	46.8	55.7	51.3	**24.6**	40.4	46.9
	Conditional DETR[8]	56.3	70.6	65.1	0.0	27.9	56.4
	DINO[10]	61.3	68.9	66.5	24.3	**49.6**	61.3
	AI-DETR-S	59.5	69.5	66.0	1.1	33.6	59.6
	AI-DETR-M	**65.2**	**73.8**	**70.5**	14.4	46.2	**65.3**

注：加粗字体表示最佳结果.

入表G2中在不同蝴蝶数据集检测精度显著的SparseRCNN作为对比方法. Open Method指Brain-Tumor数据集开源网站的最佳模型，仅提供了AP_{50}的评价指标.

表G4的实验结果显示，本文改进的AI-DETR-M在KITTI、SIMD和Butterfly_2025的AP值均达到最佳，分别优于基线模型0.7个百分点、1.1个百分点和3.9个百分点，在Brain-Tumor数据集中，与基线模型DINO相比，该模型的AP仅提升了0.2个百分点，其检测精度甚至低于单尺度目标检测方法Conditional DETR，这符合“没有免费的午餐”理论. 当然，这种情况或许是由于Brain-Tumor数据集图像质量偏低，模型训练时经插值方法放缩，对多尺度目标特征提取产生了负面影响. 例如，多尺度检测方法SparseRCNN在此数据集的AP也是在表G4所示的4个数据集中最低的. 相比之下，本文改进的AI-DETR-S模型在Brain-Tumor数据集检测性能达到最佳，其AP指标较基线模型提升了1.2个百分点，AP_{50}指标比该数据集开源的最佳方法高出了0.4个百分点，说明单尺度特征在低质图像检测中存在优势. 在SIMD和Butterfly_2025数据集上，AI-DETR-S模型的AP指标达到除DINO和AI-DETR-M外最佳结果，较基线模型分别提升1.8个百分点和3.2个百分点，较SparseRCNN模型高了9.6个百分点和12.7个百分点. AI-DETR-S在KITTI数据集的AP弱于SparseRCNN，但AP_{50}和AP_L较SparseRCNN高出1.2个百分点和5.3个百分点. 上述分析说明，本文提出的AI-DETR在处理不同领域目标检测任务时具有优势.

对比表G4各模型的AP_S和AP_M指标发现，本文提出的创新点改进的Conditional DETR，即AI-DETR-S或AI-DETR-M模型在KITTI、SIMD和Brain-Tumor数据集性能达到最佳. 这一结果说明本文方法利于提升模型对小目标的检测性能. 对Butterfly_2025数据

集,本文 AI-DETR-M 在小尺寸和中等尺寸目标的检测精度比最佳模型分别低 10.2 个百分点和 3.4 个百分点,这可能是因为模型出现了学习偏置,导致无法有效检测数据集中处在实例数量分布尾部的小尺寸目标. 表 G4 中各模型的 AP_L 指标比较揭示了基于 DETR 的方法显著高于基于卷积的 SparseRCNN 方法,说明 DETR 的全局注意力有利于检测大尺寸目标,同时也说明本文提出的 AI-DETR(AI-DETR-S 或 AI-DETR-M)模型在没有削弱 DETR 对大尺寸目标检测能力的前提下,增强了其对小尺寸和中等尺寸目标的检测能力.

表 G4 各模型对 4 种不同领域数据集的目标检测结果表明了本文模型 AI-DETR(AI-DETR-S 或 AI-DETR-M)的 AP 指标均达到最佳,说明本文提出模型 AI-DETR 对目标检测有一定的领域通用性. 此外,本文提出模型 AI-DETR 在各数据集的增益和表现不同,此外,本文提出的 AI-DETR 模型在各数据集上的增益和表现存在差异,这可能是因为数据集的先验偏置所导致的. 这一现象需要进一步深入研究.

作者简介

鲁银圆 男,1994 年 12 月出生于河南省许昌市. 现为陕西师范大学计算机科学学院博士研究生. 主要研究方向为图像处理、目标检测、深度学习等.

E-mail: zzu_luyy@163.com

许升全 男,1967 年 6 月出生于陕西省西安市. 现为陕西师范大学生命科学学院教授、博士生导师. 主要研究方向为昆虫分类与进化、进化基因组学、生物信息学等.

E-mail: xushengquan@snnu.edu.cn

谢娟英 女,1971 年 4 月出生于陕西省西安市. 现为陕西师范大学计算机科学学院教授、博士生导师,CCF 杰出会员. 主要研究方向为机器学习、数据挖掘、生物医学大数据分析、智能信息处理等.

E-mail: xiejuany@snnu.edu.cn