基于改进 DETR 的密集行人检测算法研究

宋天泽[1,2]，曹从军[1,2]，何佳琪[1,2]，王旭升[1,2]，刘晨煜[1,2]

（1.西安理工大学印刷包装与数字媒体学院，陕西 西安 710054；2.陕西省印刷包装工程技术研究中心，陕西 西安 710054）

摘　要：密集行人检测是行人检测领域的一大研究热点。针对密集行人检测场景中被遮挡目标及小目标行人易漏检的问题，提出一种改进 DETR 的目标检测算法 Pe-DETR。采用基于多头自注意力机制的 Dino-DETR 作为基准模型，因自注意力机制缺少捕获局部特征的能力，导致密集行人检测效果较差，对前馈神经网络（FNN）进行改进，设计通道注意力深度卷积前馈神经网络 DWSEFNN，使模型可以提取到更多局部细节特征。针对 ResNet50 骨干网络对重要特征提取效率较低的问题，采用 Swin Transformer-L 作为特征提取网络，提升骨干网络对重要特征的提取能力，同时使 Pe-DETR 完全基于注意力机制搭建，结构中不包含深度卷积结构。针对密集行人场景中目标数量多与 DETR 检测器中稀疏匹配的矛盾问题，应用密集不同查询有效应对行人密集的场景，且不会引入无效的相似查询。在 CrowdHuman 密集行人检测数据集上的实验结果表明，所提行人检测算法 Pe-DETR 相比 Dino-DETR 算法的平均精度（AP）@0.5 提高了 3.7 百分点，AP 提高 4.5 百分点，在密集行人检测任务中改进后 Pe-DETR 算法的准确率明显优于其他端到端模型。

关键词：行人检测；目标检测；深度卷积；迁移学习；自注意力机制

中图分类号：TP391.43　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0070106

Research on Dense Pedestrian Detection Algorithm Based on Improved DETR

SONG Tianze[1,2], CAO Congjun[1,2], HE Jiaqi[1,2], WANG Xusheng[1,2], LIU Chenyu[1,2]

(1.Faculty of Printing, Packaging Engineering and Digital Media Technology, Xi′an University of Technology, Xi′an 710054, Shaanxi, China;
2.Printing and Packaging Engineering Technology Research Center of Shaanxi Province, Xi′an 710054, Shaanxi, China)

【Abstract】Dense pedestrian detection is a research hotspot in the field of pedestrian detection. This study proposes an improved DETR target detection algorithm, Pe-DETR, to address the problem of occluded targets and small target pedestrians being prone to missed detection in dense pedestrian detection scenes. This algorithm uses Dino-DETR, which is based on the multi-head self-attention mechanism, as the benchmark model. However, the self-attention mechanism lacks the ability to capture local features, resulting in poor detection of dense pedestrians. To address this issue, this study enhances Feedforward Neural Network (FNN) and proposes channel attention convolutional feedforward neural network DWSEFNN to extract more local detailed features. In response to the low efficiency of the ResNet50 backbone network in extracting important features, Swin Transformer-L is adopted as the feature extraction network. Simultaneously, Pe-DETR is completely built based on the attention mechanism, and the architecture does not contain a deep convolution structure. To handle the contradictions between the large number of targets in dense pedestrian scenes and sparse matching in the DETR detector, densely different queries are applied to handle pedestrian-dense scenes without introducing invalid similar queries. Experimental results on the CrowdHuman dense pedestrian detection dataset show that, compared with the Dino-DETR algorithm, the proposed pedestrian detection algorithm Pe-DETR achieves an improvement of 3.7 percentage points in Average Precision (AP)@0.5 and an increase of 4.5 percentage points in AP. In dense pedestrian detection tasks, the improved Pe-DETR algorithm demonstrates significantly higher accuracy than other end-to-end models.

【Key words】pedestrian detection; object detection; depthwise convolution; transfer learning; self-attention mechanism

0　引言

行人检测[1-3]是计算机视觉领域中的重要问题，目的是快速检测出视频或图像中的行人并准确标注位置。行人检测主要包括预测、定位和标记行人的位置，以获得行人的位置和行为等信息。一个能准

基金项目：陕西省重点科研基地项目（2023HBGC-18）。

作者简介：宋天泽（CCF 学生会员），男，硕士研究生，主研方向为目标检测、图像描述；曹从军（通信作者），教授、博士；何佳琪，硕士研究生；王旭升，副教授、博士；刘晨煜，硕士研究生。

收稿日期：2024-07-11　　**修回日期：**2024-11-08　**E-mail：**caocongjun@xaut.edu.cn

确检测行人的算法在自动驾驶、智能监控和高级人机交互等应用中起着至关重要的作用。这些智能监控机器或者设备依赖于行人检测技术，能快速准确地检测到行人及其位置，只有先快速准确地检测到行人，才能进一步对人的行为、表情等进行分析并判断，从而带来最佳的人机交互体验。此外，行人检测技术是行人重识别[4-7]、人体姿态估计[8-10]和人员搜索[11-14]等研究课题的基本组成部分。

基于深度学习的行人检测方法通过深度卷积神经网络取得了显著进展，能够自动学习和提取图像特征。以 YOLO[15]系列为代表的单阶段检测器直接回归物体的类别概率和位置的坐标，通常将图像上的所有位置视为潜在目标，并尝试将每个感兴趣的区域分类为背景或目标，经过单次检测即可直接得到最终的检测结果，因此，具有更快的检测速度，大多数可以满足实时性的要求，但准确度低于双阶段检测器。以 Faster R-CNN[16]为代表的双阶段检测器首先训练候选区域生成网络，然后训练目标区域检测的网络，通过分类和回归这 2 个分支，实现对候选目标类别的判定和目标位置的确定。检测器的准确度较高，相对单阶段目标检测器的速度慢。

DETR[17]是首个基于多头注意力机制的目标检测器，基于 Transformer 的编码器-解码器架构[18]，消除了对手动组件的依赖。此外，DETR 通过匈牙利匹配在预测集合与真实目标之间建立一对一关系，并基于该匹配构建匈牙利损失进行端到端训练，简化了目标检测流程，但是 DETR 也存在一些问题。首先，DETR 需要 500 个训练周期才能达到与 Faster R-CNN 检测器相同的准确率。此外，DETR 的查询是稀疏的，在密集场景中不能充分发挥模型性能。

在行人检测中，行人在视频或者图片中通常因遮挡问题而呈现不完整状态，比如其他物体对行人的遮挡或者行人个体之间相互遮挡。这一现象导致在训练和检测的过程中，输入检测器的目标并不完整，使得模型可能误将背景环境或其他行人的局部特征作为学习目标，从而对检测器产生干扰，因此对于密集遮挡的行人检测问题有待进一步研究。本文以 Dino-DETR[19]为基线，对骨干网络、前馈神经网络(FNN)进行改进，并引入一种密集不同查询的方式提升密集场景下行人检测的准确性。

本文提出一种改进的目标检测算法 Pe-DETR，使用 Swin Transformer-L 骨干网络进行迁移学习，以解决基于多头自注意力机制模型训练过程中的网络性能下降、梯度消失、梯度爆炸、训练开销大等问题。在密集场景中因行人相互重叠、遮挡导致目标不完整，稀疏查询不能发挥模型性能，本文利用密集不同查询模块提高检测准确性，密集查询能解决密集目标与稀疏查询之间的矛盾，而引入不同查询可降低密集查询之间的相似性，从而缓解其在匈牙利匹配阶段引发的匹配歧义和优化问题。此外，引入了通道注意力深度卷积前馈神经网络 DWSEFNN 模块，在原有 FNN 中加入深度卷积与 SE(Squeeze-and-Excitation)通道注意力机制，以增强模型对 Transformer 自注意力机制中图像细节特征的提取能力，并且保持了前向传播的尺寸不变，从而提高了模型在密集场景下捕获行人关键性细节特征的能力。

1 基于改进 DETR 的密集行人检测器

本文研究密集行人检测和非常规姿态行人检测，为了提高网络在密集场景下的检测能力，提出一种基于改进 DETR 的行人检测网络 Pe-DETR，网络结构如图 1 所示(彩色效果见《计算机工程》官网 HTML 版，下同)。

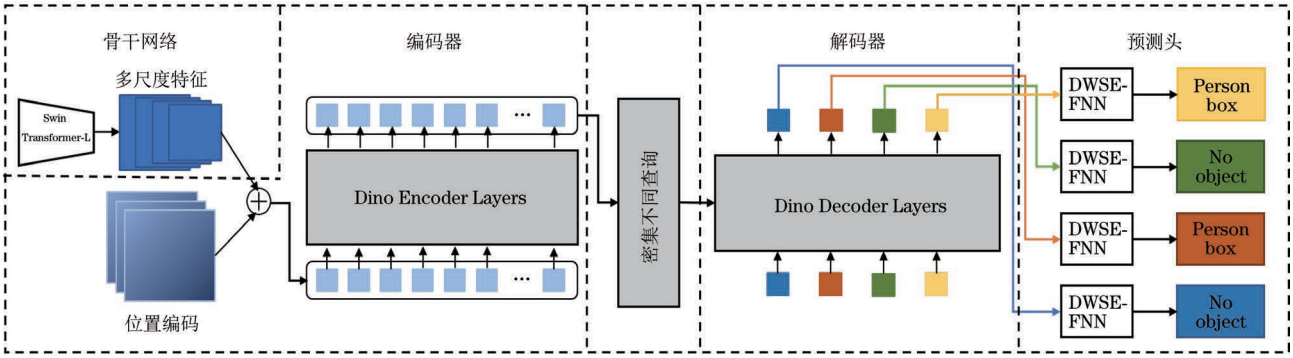

图 1 Pe-DETR 网络结构

Fig. 1 Structure of Pe-DETR network

该网络采用 Dino-DETR 作为主要框架，在特征提取阶段使用 Swin Transformer[20]作为骨干网络，使用 ImageNet22K 上的预训练模型进行迁移学习。采用密集不同查询模块，密集查询通过平铺平

移不变的查询实现，不同查询通过 1 个 0.7 阈值的非极大值抑制(NMS)[21]筛选实现。本文提出替换原模型中的 FNN，DWSEFNN 中的深度卷积可以提取图像中的细节特征，作为 1 个即插即用的模块，适用于所有 Transformer 结构的网络模型。

1.1 基于 Swin Transformer 骨干网络的迁移学习

1.1.1 Swin Transformer-L 骨干网络

本文基于 Swin Transformer-L 的核心思想是通过分层的、分区域的注意力机制来处理图像，其骨干网络架构如图 2 所示，主要由图像块拆分层、堆叠层、层归一化、全局池化层以及全连接层组成。H 为输入图像的高度，单位为像素；W 为输入图像的宽度，单位为像素；C 为特征图维度；LN 为归一化层；W-MSA 为窗口多头自注意力结构；SW-MSA 为移动窗口多头自注意力结构；MLP 为多层感知器；stage1～4 为堆叠模块；Swin Transformer block 为由归一化层、窗口多头自注意力结构、多层感知器和移动窗口多头自注意力结构组成的模块。整个模型采取层次化的设计，一共包含 4 个堆叠层，每个堆叠层都会将输入特征图的分辨率缩小，像卷积神经网络一样逐层扩大感受野。首先将 $H\times W\times 3$ 的 RGB 三通道图像输入到图像块拆分层中进行分块，每 4×4 的像素分成 1 个图像块，通道数 $C=3$，通过图像块拆分层在通道方向展平后，特征图像维度变为 48，即图像形态由[$H\times W\times 3$]变成了[$H/4\times W/4\times 48$]；然后经过 4 个堆叠层构建不同大小的特征图，每个堆叠层都会缩小输入特征图的分辨率。在 Swin Transformer-L 中，第 1 个堆叠层包含 1 个线性嵌入层和 2 个 Swin Transformer block，第 2 个和第 4 个堆叠层由 1 个图像块合并层和 2 个 Swin Transformer block 组成，第 3 个堆叠层包含 1 个图像块合并层和 18 个 Swin Transformer block。传统的 Transformer 是基于全局多头自注意力机制构建的，划分 8 个头分别计算，然后将 8 个头的计算结果进行融合得到多头自注意力机制的计算结果，公式如下：

$$h_{\text{head}_i}=A_{\text{Attention}}(\boldsymbol{QW}_{iQ},\boldsymbol{KW}_{iK},\boldsymbol{VW}_{iV}),\ i=1,2,\cdots,8 \tag{1}$$

$$M_{\text{MultiHead}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(h_{\text{head}_1},h_{\text{head}_2},\cdots,h_{\text{head}_8})\boldsymbol{W}^0 \tag{2}$$

式中：$\boldsymbol{W}_{iQ}$、$\boldsymbol{W}_{iK}$、$\boldsymbol{W}_{iV}$ 是 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 各自生成的矩阵。单头自注意力机制的计算公式如下：

$$A_{\text{Attention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{QK}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

式中：$\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 是矩阵；$\boldsymbol{QK}^{\text{T}}$ 乘积用于计算查询与键之间的相似度，计算结果用于确定模型将多少注意力放在这个查询上；d_k 是 $\boldsymbol{K}$ 的维度，用来对点积进行缩放；Softmax 函数用来将数值向量归一化为概率分布向量，使它们满足概率分布的性质，即概率之和为 1。

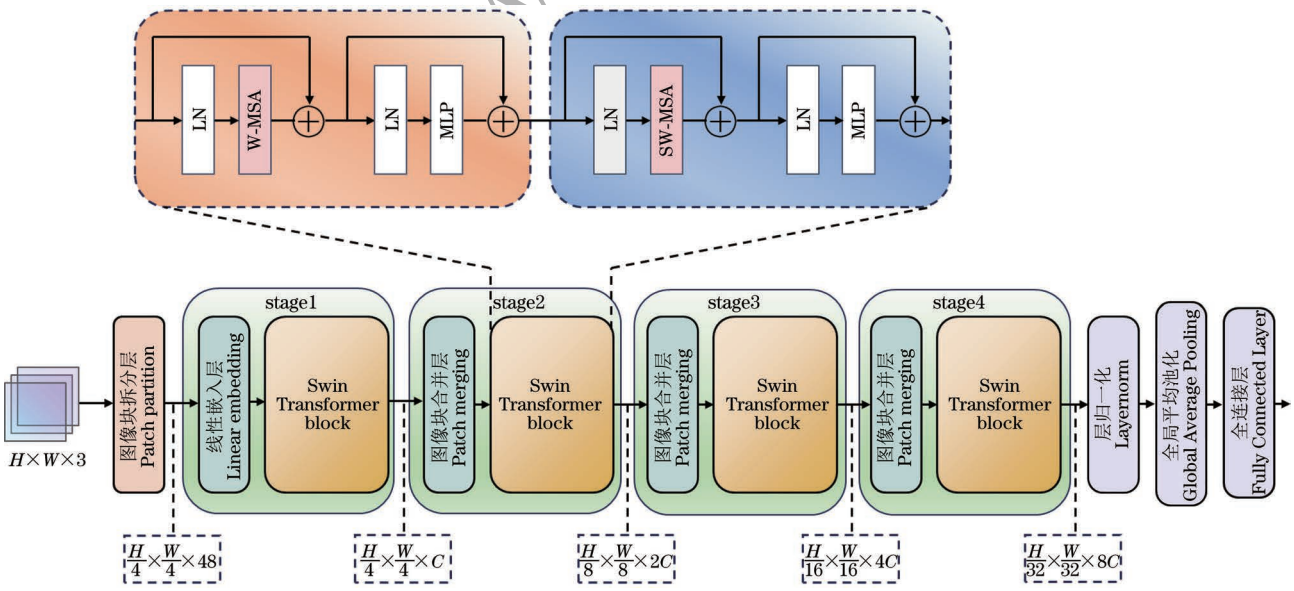

图 2 Swin Transformer-L 结构

Fig. 2 Structure of Swin Transformer-L

W-MSA 将输入图像分割为多个小窗口，每个操作可以将多头注意力机制操作限制在分割后的窗口内，一方面引入卷积的局部性，另一方面可以减少计算量。图像中的不同区域之间可能存在局部相关性，为了使各窗口之间能更好地进行信息交互，SW-MSA 在 W-MSA 的基础上加入了滑动窗口操作，目的是在图像中平移并拼接窗口以实现不同窗口之间的信息交互。滑动窗口的过程如

图 3 所示。窗口滑动后会导致各窗口的大小不一致，为保证计算过程的一致性，通过对不同大小的窗口进行拼接后再计算每个窗口的注意力。W-MSA 与 SW-MSA 在网络中成对使用，并与线性层、前向传播网络、残差连接结构共同构成 Swin Transformer block。

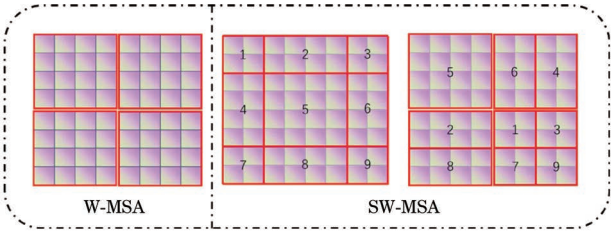

图 3 移动窗口多头自注意力过程

Fig. 3 Process of the shifted window multi-head self-attention

1.1.2 迁移学习

迁移学习的原理示意图如图 4 所示。对于给定的目标域，借助已有源域和源任务的知识，建立从目标域数据到标签的映射函数，完成目标任务。在实际应用中密集行人数据集中目标个数多、遮挡面积大，导致大规模手工标注要消耗大量人力，模型训练开销大、耗时长、难收敛。本文提出一种迁移学习方法，将源域上学习的目标检测知识迁移至密集行人检测领域，利用源域中预先训练好的部分网络结构和连接参数。采用预训练微调的迁移学习方法，利用 ImageNet22K 超大规模数据集训练好的网络模型初始化骨干网络参数，先在 COCO 数据集上进行 12 次迭代预训练，并利用目标域行人数据集微调模型，因目标域所检测的行人类别属于源域，且数据量远远小于预训练数据集，所以训练时不冻结骨干网络，只将分类头中的类别数更改为 1 类，即“行人”类，随后冻结第 1 层在目标域密集行人数据集上微调模型，在此基础上更新权重参数，避免因数据不足导致的过拟合，并解决 Swin Transformer 训练难收敛的问题。

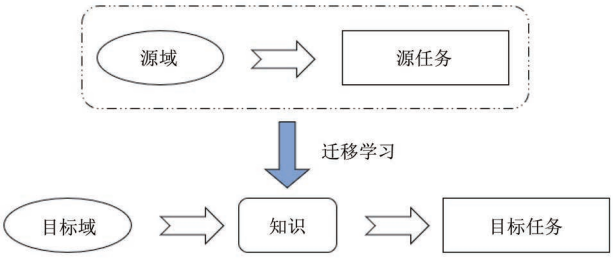

图 4 迁移学习原理示意图

Fig. 4 Schematic diagram of transfer learning principles

1.2 通道注意力深度卷积前馈神经网络

本文提出的通道注意力深度卷积前馈神经网络由全连接层、深度卷积、高斯误差线性单元激活函数和通道注意力机制构成。不同于常规卷积操作，深度卷积中的 1 个卷积核只负责 1 个通道，卷积核的数量与输入图像的通道数相等，并且不改变输出的通道数。GELU 激活函数与 ReLU 激活函数相比，具有更平滑的非线性特征，有助于提高 Transformer 模型的性能。

FNN 存在于 DETR 的编码器、解码器和输出头中。DETR 是基于多头自注意力机制的目标检测器，注意力机制就是对特征向量加权求平均的过程，即线性变化。在实际任务中，如果利用非线性变化来学习更复杂的非线性特征、拟合真实的数据，就需要 FNN 的表达能力。DETR 中的 FNN 由 2 个全连接层和 ReLU 激活函数构成，由于只对特征图进行了 1×1 的卷积，因此缺乏局部区域内的信息交互。HOWARD 等[22]提出基于深度可分离卷积的 MobileNet 架构，在 MobileNetV2 中首次引入反向残差块[23]，由一系列 1×1 点卷积和 3×3 深度卷积组成，与 Transformer 中 FNN 的唯一区别是 MobileNetV2 在反向残差块中加入了 1 个深度卷积，各通道经过 3×3 的深度卷积核重新聚合为新的特征。在高光谱图像恢复任务 DSTrans[24]中提出了双流前馈网络，用于提取并行支路中的全局信息和局部细节信息。在骨干网络 localViT[25]的研究中通过在 FNN 中引入深度卷积来增加 Vision Transformer 的局部性，并将相同的局部性机制成功应用于 4 种基于多头自注意力机制的骨干网络中。文献[26]在骨干网络 TransNeXt 的研究中提出在 GLU 门控分支的激活函数前加入深度卷积，可以将其转化为基于最近邻特征的门控通道注意力机制，同时提到在 Vision Transformer 中引入深度卷积，可以被视为加入条件位置编码(CPE)的一种形式[27]。

受上述研究启发，为了将局部性有效引入到检测器中来提升密集场景下行人检测精度，将深度卷积与 FNN 相结合并设计多种不同的结构，如图 5 所示。经过实验对比，本文所提的通道注意力深度卷积 FNN 取得了最好的效果，在 FNN 的第 1 个全连接层后进行深度卷积运算，并使用 GELU 激活函数进行非线性变换，与 ReLU 激活函数相比，GELU 更平滑，避免了 ReLU 将输入裁剪到 0 时可能产生梯度消失的问题，在计算机视觉任务中表现更好。GELU 激活函数的公式如下：

$$\mathrm{GELU}(x)=xP(X\leqslant x)=x\phi(x) \tag{4}$$

式中：$\phi(x)$表示二项式分布。

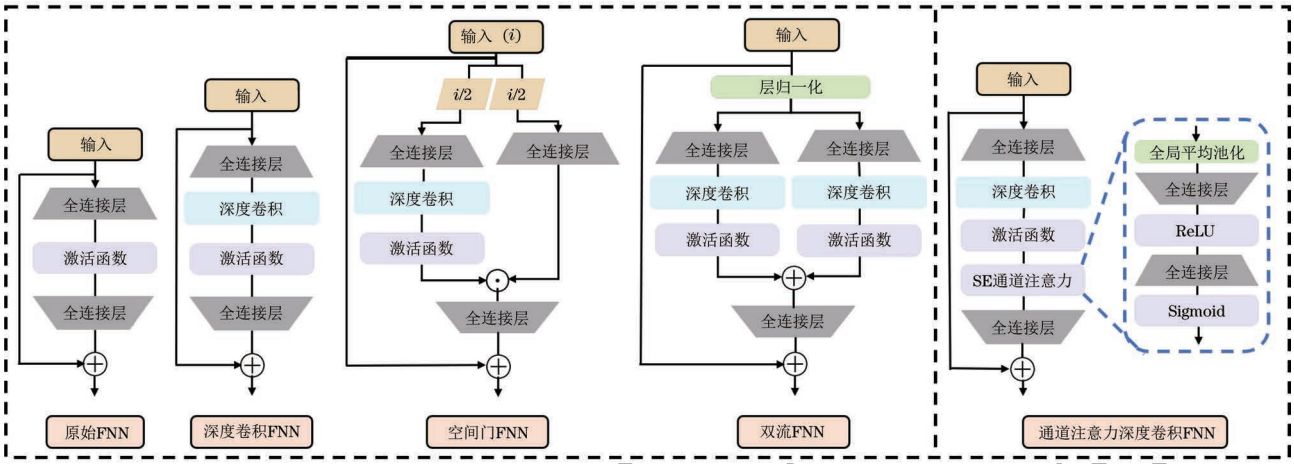

图 5 常见前馈神经网络与通道注意力深度卷积前馈神经网络结构

Fig. 5 Structures of common feedforward neural networks and channel attention depthwise convolutional feedforward neural network

由于式(4)中二项式分布无法直接计算,因此在实际应用中使用近似式替代计算,具体表达式如下:

$$\mathrm{GELU}(x)=0.5x\left(1+\tanh\left[\sqrt{\frac{2}{\pi}}(x+0.044\ 715x^3)\right]\right) \tag{5}$$

深度卷积只按通道进行计算,虽然解决了因常规卷积带来模型计算量急剧增加的问题,但无法关注到通道之间的相关信息,因此加入 SE 通道注意力机制,给予各通道不同的权重。SE 模块主要包含压缩和激励,压缩部分对特征进行全局平均池化,激励部分通过 2 个全连接层构建通道之间的相关性来生成 1 个权重值,最后把得到的注意力赋予到每个通道特征上。DWSEFNN 的计算公式为:

$$\mathrm{DWSEFNN}(x)=\mathrm{Linear}(\mathrm{SE}(\mathrm{GELU}(\mathrm{DW}(\mathrm{Linear}(x))))) \tag{6}$$

式中:DW 表示深度卷积运算;GELU 表示高斯误差线性单元激活函数;SE 为通道注意力机制。通道注意力机制的计算公式如下:

$$z=f_{\mathrm{sq}}(x)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_{i,j} \tag{7}$$

$$g=f_{\mathrm{ex}}(z)=\sigma(W_2\delta(W_1z)) \tag{8}$$

$$y=f_{\mathrm{scale}}(x,g)=g\cdot x \tag{9}$$

式中:x 是输入特征图;z 是压缩后的特征图;g 是激励后的特征图;f_{sq} 是压缩操作;f_{ex} 是激励操作;f_{scale} 是将权重矩阵与特征图相乘;δ 和 σ 分别表示 ReLU 激活函数和 Sigmoid 激活函数。

1.3 密集不同查询模块

当前的端到端检测器要么使用密集查询,要么使用稀疏查询,这两种查询在训练过程中都无法应用于密集行人检测任务中。具体来说,稀疏查询的召回率很低,而密集查询不可避免会带来大量相似的查询,而这些相似查询会导致大量的图形处理器(GPU)内存被占用,在优化方面存在问题。在本任务中,首先像传统检测器一样设置预定义的密集查询,然后利用卷积和线性网络结构对查询进行快速筛选,最后选择不同的查询进行一对一的匹配。密集不同查询的结构如图 6 所示,它结合了传统的基于深度学习目标检测器和端到端目标检测器的优点。本研究直接根据分类分数选择前 1 350 个查询作为解码器中的密集查询,为方便比较,保留了 Dino-DETR 中的对比去噪训练,并将不同的查询数量设为 900 个。

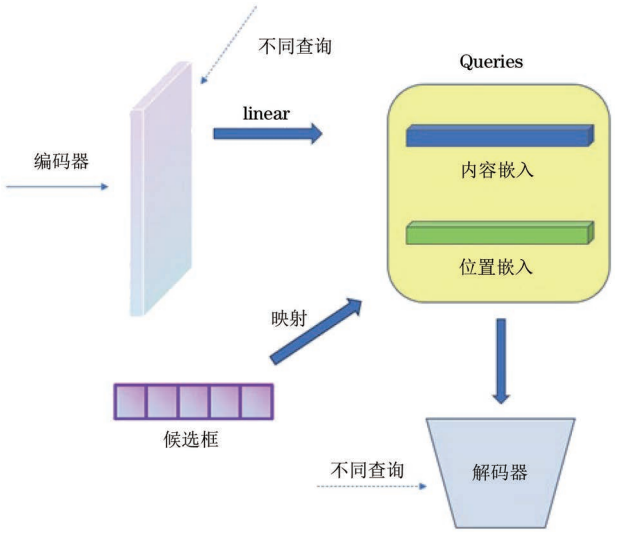

图 6 密集不同查询示意图

Fig. 6 Schematic diagram of dense distinct queries

密集不同查询没有采用 DETR 中可学习的位置嵌入,而是根据双阶段目标检测器的方法直接将每个特征映射上的特征点作为密集分布的初始查询,但每个查询的计算量都很大,密集查询会导致显存成本急剧增加。为了将计算量保持在合理范围,使用轻量级卷积以滑动窗口的方式先对所有查询进行处理,再进行之后的筛选操作。

在实现密集查询的同时不可避免会为每个行人目标带来多个查询，这些查询不会帮助模型更快收敛，甚至会导致负训练。ZHANG 等[28]提出重复的查询可以减少梯度消失现象，这极大地抑制了收敛性，一旦能够确保所选择的查询是不同的，检测器的性能可以随着查询数量的增加而不断提高。本任务中应用 1 个阈值设定为 0.7 且类无关的 NMS 来一对一分配筛选出不同的查询，只对 NMS 筛选出的不同查询计算损失。密集查询产生的过程中在同一目标附近会产生大量的候选框，这些候选框是冗余的，利用 NMS 找到单个行人目标的最佳候选框，可以达到不同查询的目标。

密集不同查询的损失函数包含主损失和辅助损失两部分，主损失与 Dino-DETR 保持一致，并在一对一匹配中使用相同的参数。密集不同查询过程中会产生大量的相似查询，这些查询不会反向传播帮助模型收敛。为了更充分利用这些查询，在计算主损失之前先进行辅助损失计算，辅助损失遵循 TOOD 算法的设计，对密集查询采用一对多赋值，允许密集梯度和更多正样本来加速模型收敛，选择每个真实目标损失最小的 8 个样本作为正样本，使用 Quality Focal Loss 计算分类损失，在使用 GIoU 损失函数计算每个正样本的回归损失之前根据分类的目标更新权重。

2 实验与结果分析

2.1 实验环境及参数设置

本文的实验在 MMDetection3.0 环境中进行，MMDetection 是 1 个基于 PyTorch 的目标检测开源工具箱。计算平台为装有 4 张 NVIDIA GeForce 2080 GPU 卡的工作站，操作系统为 Windows 10。输入的 Batch Size 设为 1，单个实验训练的 Epoch 设为 12，18 000 次线性学习率预热，网络的学习率设为 5×10^{-5}，在第 9 个和第 11 个 Epoch 结束后对学习率进行 MultiStepLR 衰减，$\gamma=0.1$。

2.2 实验所用数据集

CrowdHuman 的训练集、验证集和测试集分别包括 15 000、4 370 和 5 000 幅图像。图片上的人体实例包含人体可见区域边界框标注、头部区域边界框标注和人体整体边界框标注。这是第 1 个专门针对人群问题的数据集，1 个图像中行人的平均数量为 22.6，2 个行人实例之间的成对重叠（大于 0.5 IoU）的平均数量是 2.4，两项指标都比原有的数据集要大得多。本文为统一不同检测器上的实验结果，将数据集转化为 COCO 的标注格式进行训练。

Human-art 包含 5 个现实和 15 个虚拟场景的 50 000 张图像，是支持行人检测、关键点检测及文本描述的多场景大规模数据集，包括杂技场景、电影场景、戏剧场景、舞蹈场景和角色扮演场景，每类数据训练集和验证集包含 1 750 张和 250 张图像。Human-art 中的图像均以人体为核心，支持对姿态、装扮更具包容性的行人检测器训练。在本文实验中使用 5 类现实场景中的图像测试模型检测不同场景下行人装扮的性能。

2.3 评价标准

在待检测图像中，检测目标作为正样本，则有与之对应的无关负样本，因此目标检测算法的预测结果共包含 4 种：预测为正样本的正样本，预测为负样本的负样本，预测为负样本的正样本，预测为正样本的负样本。精准率（P）又称查准率，用于表示在预测结果中，预测正确的目标数量占所有被检测目标数量的比例，可以衡量算法检测的准确度，计算公式为：

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}}$$

式中：N_{TP} 表示实际为正被预测为正的样本数量；N_{FP} 表示实际为负但被预测为正的样本数量。

召回率（R）又称查全率，用于表示在预测结果中，预测为正样本的目标数量占所有实际正样本的比例，可以衡量算法识别正负样本的能力，计算公式为：

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}}$$

式中：N_{FN} 表示实际为正但被预测为负的样本数量。

P-R 曲线是以 R 为横轴，P 为纵轴。P-R 曲线与坐标轴所围面积称为平均精度（AP）。本文中的 AP 值为 10 个 IoU 阈值 0.50∶0.05∶0.95 上取平均值，AP@0.5 是在 1 个单一 0.50 的 IoU 上计算得到的。超过均值的 IoU 使探测器更好定位。APS 是对于图像中分辨率较小目标的平均检测精度，APM 是对于图像中分辨率中等目标的平均检测精度，APL 是对于图像中分辨率较大目标的平均检测精度。其中较小目标的像素面积小于等于 32×32，中等尺寸目标的像素面积范围为（32×32）～（96×96），将像素面积大于 96×96 的归为大尺寸目标。平均精度均值（mAP）是所有类别的平均值。在本研究中由于只检测人一个类别，因此没有区分 AP 和 mAP。

2.4 基于 CrowdHuman 数据集的实验结果分析

在 CrowdHuman 数据集上验证所提方法的优越性，与现有先进端到端检测器相比，为保证公平，所有检测器均在 MSCOCO 数据集上进行预训练再迁移至 CrowdHuman 数据集，使用 AP、AP@0.5、AP@0.75、APS、APM、APL 作为模型性能的衡量指标。对比实验选择的检测器均为使用 4 级多尺度特征的端到端检测器，因此未选择只使用单一尺度特征的 DETR 检测器。实验结果如表 1 所示。从表 1 可以看出，使用 Swin Transformer-L 骨干网络进行迁移学习的 Pe-DETR 相较于其他端到端目标检测器对密集场景下的行人检测更加准确，与 Deformable-DETR、Dab-DETR、Condiontional-DETR 相比，Pe-DETR 可以用更少的训练次数得到更高的准确性，与 Dino-DETR 相比，本模型检测各尺寸行人目标时的准确性均有不同程度提升，特别是对于小目标检测的准确率提升最多，同时改进后的模型在训练中的损失更小，收敛得更快。

表 1 在 CrowdHuman 数据集上端到端检测模型的性能对比

Table 1 Comparison of the performance of an end-to-end detection model on CrowdHuman dataset

模型	Backbone	Epoch/次	AP	AP@0.5	APS	APM	APL
Deformable-DETR	ResNet50	50	0.448	0.783	0.199	0.388	0.490
Dab-DETR	ResNet50	50	0.374	0.754	0.142	0.342	0.453
Condiontional-DETR	ResNet50	50	0.342	0.700	0.083	0.266	0.440
Dino-DETR	ResNet50	12	0.475	0.836	0.325	0.479	0.534
Pe-DETR(Ours)	Swin Transformer-L	6	0.513	0.861	0.377	0.528	0.565
Pe-DETR(Ours)	Swin Transformer-L	12	0.520	0.873	0.395	0.550	0.571

为了验证本节所采用的训练方法与 Swin Transformer-L 作为骨干网络的作用，在 CrowdHuman 数据集上进行消融实验，分别使用 ResNet50 和 Swin Transformer-L 骨干网络进行正常训练和迁移学习训练，实验结果如表 2 所示。在 CrowdHuman 数据集上的消融实验结果证明了 Swin Transformer-L 骨干网络对于密集行人目标的特征提取性能优于 ResNet50，同时也证明了迁移学习在密集行人检测任务中的效果，在使用迁移学习后的 ResNet50 与 Swin Transformer-L 骨干网络分别将 AP@0.5 提升了 15.1 与 14.6 百分点。ResNet50 在迁移学习后小目标检测性能下降，这可能是源域与目标域数据集的差异所导致的，而 Swin Transformer-L 将小目标检测的平均精度提升了 2.4 百分点，证明了通过大规模数据的预训练，该骨干网络迁移至下游任务中的效果更好。本方法的提出为端到端密集行人检测器提供了一种新的范式。

表 2 在 CrowdHuman 数据集上消融实验结果

Table 2 Ablation experimental results on CrowdHuman dataset

骨干网络	迁移学习	AP@0.5	APS	APM	APL
ResNet50	×	0.685	0.356	0.315	0.408
ResNet50	√	0.836	0.325	0.479	0.534
Swin Transformer-L	×	0.723	0.371	0.386	0.453
Swin Transformer-L	√	0.869	0.395	0.538	0.561

为了证明本文所提算法的整体性能，将其与现有的经典目标检测算法，包括 YOLOv5、YOLOv7 进行对比分析，所有实验都是在 CrowdHuman 数据集上进行的，实验结果如表 3 所示。从表 3 可以看出，虽然改进后的模型推理速度有所下降，但由于引入了注意力机制，模型的 AP@0.5 显著优于 YOLOv5 等经典算法，对小目标行人检测时的准确性较好。

表 3 在 CrowdHuman 数据集上不同算法的对比结果

Table 3 Comparison results among different algorithms on CrowdHuman dataset

算法	Backbone	AP@0.5	FPS
YOLOv5	YOLOv5n	0.812	68.5
YOLOv7	YOLOv7-tiny	0.831	93.9
Deformable-DETR	ResNet50	0.783	25.1
Condiontional-DETR	ResNet50	0.700	30.8
Pe-DETR(Ours)	Swin Transformer-L	0.873	28.6

为验证本文所提即插即用模块 DWSEFNN 比其他不同结构的 FNN 更有效，用图 5 所示的结构分别替换原始 FNN，其余部分保持不变。所有对比实验均以 Dino-DETR 作为基础，以 ResNet50 作为骨干网络，在 MSCOCO 数据集上进行预训练再迁移至 CrowdHuman 数据集，结果如表 4 所示。在加入深度卷积后，不同结构的 FNN 与未加入前原始 FNN 相比，对密集行人检测的准确性均有不同程度的提高，其中直接在全连接层后加入深度卷积的 FNN 比结构更复杂的双流 FNN 和空间门 FNN 效

果更好。与原始 FNN 相比，加入深度卷积的 FNN 将 AP 提高了 1.1 百分点，AP@0.5 提高了 0.5 百分点，证明了利用深度卷积能够为基于注意力机制的行人检测器引入局部性。用包含通道注意力模块的 DWSEFNN 替换原始 FNN，在不引入额外计算量的前提下，与原始 FNN 相比，AP 提高了 1.4 百分点，AP@0.5 提高了 0.7 百分点。

表 4　不同结构 FNN 的实验结果

Table 4　Experimental results of FNN with different structures

方法	AP	AP@0.5
FNN	0.475	0.836
深度卷积 FNN	0.486	0.841
双流 FNN	0.481	0.834
空间门 FNN	0.480	0.837
DWSEFNN	0.489	0.843

在 CrowdHuman 数据集上的实际检测结果示例如图 7 所示。从图 7 可以看出，即使图片中的行人目标较为拥挤，存在不同程度的遮挡，且年龄、装扮、姿态各异，使用本文所提的 Pe-DETR 均能得到较高的得分和高质量的预测边界框，从而完成检测。

图 7　Pe-DETR 检测结果示例

Fig.7　Example of Pe-DETR's detection results

2.5　基于 Human-art 数据集的实验结果分析

为了进一步测试本文所提方法在不同姿态、装扮行人上的效果，在 Human-art 数据集的 5 类真实场景下进行实验，实验结果如表 5 所示。从表 5 可以看出，在 Human-art 数据集上的实验结果与 CrowdHuman 数据集上的对比结果相似，在 Human-art 的 5 类真实场景数据集上也取得了出色的效果，证明了 Pe-DETR 能准确检测出不同姿态、装扮的行人。

表 5　在 Human-art 数据集上的实验结果

Table 5　Experimental results on Human-art dataset

类别	mAP	AP@0.5	AP@0.75	APM	APL
杂技	0.786	0.958	0.892	0.501	0.894
角色扮演	0.740	0.952	0.861	0.364	0.859
舞蹈	0.692	0.908	0.791	0.106	0.774
戏剧	0.641	0.896	0.733	0.472	0.642
电影	0.553	0.837	0.623	0.097	0.580

2.6　消融实验

为验证所提各改进模块的有效性，本文设置了消融实验，实验结果如表 6 所示。在 CrowdHuman 数据集上以 AP、AP@0.5 两个主要评价指标进行对比。实验 A 用 DWSEFNN 替换原始 FNN；实验 B 在 A 的基础上加入密集不同查询；实验 C 在实验 B 的基础上用 Swin Transformer-L 作为骨干网络提取特征。从表 6 可以看出，每个改进模块都提升了密集场景下行人检测的准确性，其中实验 C 使用 Swin Transformer-L 骨干网络提取特征，为检测器带来的增益最为明显，相比实验 B，将 AP@0.5 提升了 2.6 百分点，体现了基于自注意力机制骨干网络的优越性能。

表 6　在 CrowdHuman 数据集上的消融实验结果

Table 6　Ablation experimental results on CrowdHuman dataset

编号	DWSEFNN	密集不同查询	Swin Transformer-L	AP	AP@0.5
Dino-DETR (baseline)	×	×	×	0.475	0.836
实验 A	√	×	×	0.488	0.842
实验 B	√	√	×	0.492	0.847
实验 C	√	√	√	0.520	0.873

3　结束语

针对密集场景下行人目标数量多、遮挡面积大的问题，本文提出一个完全基于自注意力机制的行人检测器 Pe-DETR，用于增强 DETR 在行人检测方面的性能。首先，采用 Swin Transformer 作为特征提取网络，并用迁移学习的方式使模型更快收敛。引入密集不同查询模块，提升了模型对密集和被遮挡目标的检测能力。设计即插即用的 DWSEFNN 模块来取代 FNN，补充了基于自注意力机制模型欠缺的局部细节特征。实验结果表明，在 CrowdHuman 数据集上进行验证，Pe-DETR 的 AP 达到了 0.52，AP@0.5 达到了 0.873。在各个数据集上的消融实验验证了不同模

块的有效性，并对不同结构的 FNN 进行对比，验证出实验效果最佳的网络框架。虽然采用注意力检测结构的算法在处理密集场景和遮挡问题上展现了显著的性能优势，但检测速度无法和单阶段目标检测模型相比，使用此类模型需要庞大的显存，因此下一步将降低训练所需显存并提升检测速度，满足实际场景下行人检测的需求，包括但不限于交通监控、智能安防和智慧城市等领域。

参考文献

[1] 宋晓琳. 基于深度学习的行人检测算法研究[D]. 北京：北京邮电大学，2023.
SONG X L. Research on pedestrian detection based on deep learning[D]. Beijing: Beijing University of Posts and Telecommunications, 2023. (in Chinese)

[2] CHEN N, LI M L, YUAN H, et al. Survey of pedestrian detection with occlusion[J]. Complex & Intelligent Systems, 2021, 7(1): 577-587.

[3] 张宏扬. 基于深度学习的遮挡行人检测研究[J]. 信息技术与信息化，2023(6)：217-220.
ZHANG H Y. Research on occlusion pedestrian detection based on deep learning[J]. Information Technology & Informatization, 2023(6): 217-220. (in Chinese)

[4] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2005: 886-893.

[5] VIOLA P, JONES M J. Robust real-time face detection[J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[6] 葛斌，许诺，夏晨星，等. 四流输入引导的特征互补可见光—红外行人重识别[J]. 光电工程，2024，51(9)：240119.
GE B, XU N, XIA C X, et al. Quadruple-stream input-guided feature complementary visible-infrared person re-identification[J]. Opto-Electronic Engineering, 2024, 51(9): 240119. (in Chinese)

[7] LIENHART R, MAYDT J. An extended set of Haar-like features for rapid object detection[C]// Proceedings of International Conference on Image Processing. Washington D. C., USA: IEEE Press, 2002: 1-10.

[8] BAY H, TUYTELAARS T, GOOL L V. SURF: speeded up robust features[C]//Proceedings of the 9th European Conference on Computer Vision. Berlin, Germany: Springer, 2006: 52-60.

[9] HEARST M A, DUMAIS S T, OSUNA E, et al. Support vector machines[J]. IEEE Intelligent Systems and Their Applications, 1998, 13(4): 18-28.

[10] OPITZ D, MACLIN R. Popular ensemble methods: an empirical study[J]. Journal of Artificial Intelligence Research, 1999, 11: 169-198.

[11] FREUND Y. Experiments with a new boosting algorithm[C]// Proceeding of International Conference on Machine Learning. [S. l.]: AAAI Press, 1996: 20-29.

[12] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[13] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[14] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.

[15] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 779-788.

[16] 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.

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2017: 6000-6010.

[18] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2024-06-05]. https://arxiv.org/pdf/1810.04805.

[19] ZHANG H, LI F, LIU S L, et al. Dino: DETR with improved denoising anchor boxes for end-to-end object detection[EB/OL]. [2024-06-05]. https://arxiv.org/pdf/2203.03605.

[20] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2022: 9992-10002.

[21] NEUBECK A, VAN GOOL L. Efficient non-maximum suppression[C]// Proceedings of the 18th International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2006: 850-855.

[22] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. [2024-06-05]. https://arxiv.org/pdf/1704.04861.

[23] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 4510-4520.

[24] YU D B, LI Q W, WANG X L, et al. DSTrans: dual-stream transformer for hyperspectral image restoration[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Washington D. C., USA: IEEE Press, 2023: 3728-3738.

[25] LI Y W, ZHANG K, CAO J Z, et al. LocalViT: analyzing locality in vision transformers[C]//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Washington D. C., USA: IEEE Press, 2023: 9598-9605.

[26] SHI D. TransNeXt: robust foveal visual perception for vision transformers[EB/OL]. [2024-06-05]. https://arxiv.org/pdf/2311.17132.

[27] CHU X X, TIAN Z, ZHANG B, et al. Conditional positional encodings for vision transformers[EB/OL]. [2024-06-05]. https://arxiv.org/abs/2102.10882.

[28] ZHANG S L, WANG X J, WANG J Q, et al. Dense distinct query for end-to-end object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2023: 7329-7338.

文字编辑 薛晋栋
栏目编辑 赖玉玲