中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-17

论文引用格式:Liu Tao, OuYang Hui, Gao Yimeng. XXXX. Frequency-Enhanced and Boundary-Aware Real-Time Pedestrian Detection for Autonomous Driving. Journal of Image and Graphics, XX(XX):0001-0017(刘涛, 欧阳晖, 高一萌. XXXX. 频域增强与边界感知的自动驾驶实时行人检测. 中国图象图形学报, XX(XX):0001-0017)[DOI:10.11834/jig.250437]

频域增强与边界感知的自动驾驶实时行人检测

刘涛[1,2],欧阳晖[1*],高一萌[3]

1. 辽宁工程技术大学软件学院,葫芦岛　125105; 2. 辽宁工程技术大学基础教学部,葫芦岛　125105; 3. 辽宁工程技术大学电子与信息工程学院,葫芦岛　125105

摘　要:**目的**　在自动驾驶场景中,行人尺度变化剧烈、遮挡现象频繁和复杂环境干扰等问题会导致检测性能显著下降。针对这些挑战,本文提出了一种结合多重频域增强策略与边界感知机制的实时行人检测算法FEBA-DETR(Frequency-Enhanced and Bound-Aware Detection Transformer)。**方法**　FEBA-DETR基于RT-DETR架构进行改进,通过使用频域增强、边界感知机制以及优化损失函数,提升对小目标的检测能力和遮挡情况下的检测能力。最后再结合频域子带数据增强方法,进一步提升算法在雨、雾、雪和低光照等复杂环境下的检测性能。**结果**　与原RT-DETR算法相比,FEBA-DETR在CityPerosns数据集上,AP50与AP50:95分别提升2.3%和2%,引入频域子带增强后分别提升3%和2.4%;在代表部分遮挡和严重遮挡的场景中,MR^{-2}分别下降4.92%和2.82%。**结论**　FEBA-DETR算法提升了在自动驾驶场景中对远距离小尺度行人和严重遮挡行人的检测能力,并在复杂环境下也表现出更强的鲁棒性,其性能优势在多组对比实验中得到了充分验证,能有效应对自动驾驶中的行人检测挑战,为自动驾驶系统的安全提供支持。

关键词:自动驾驶;行人检测;小目标检测;频域增强;边界感知;数据增强

Frequency-Enhanced and Boundary-Aware Real-Time Pedestrian Detection for Autonomous Driving

Liu Tao[1,2], OuYang Hui[1*], Gao Yimeng[3]

1. School of Software, Liaoning Technical University, Huludao 125105, China; 2. Department of Basic Teaching, Liaoning Technical University, Huludao 125105, China; 3. School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China

Abstract: **Objective** In autonomous driving scenarios, pedestrian detection performance can degrade significantly due to dramatic variations in pedestrian scale, frequent occlusions, and interference caused by complex backgrounds. To overcome these difficulties, we introduce a novel real-time pedestrian detection algorithm named FEBA-DETR (Frequency-Enhanced and Bound-Aware Detection Transformer), which integrates multiple frequency-domain enhancement strategies with a boundary-aware mechanism. By leveraging these complementary components, the proposed method is designed to achieve more reliable pedestrian detection under the aforementioned challenging conditions. **Method** The Real-Time Detection Transformer(RT-DETR) has demonstrated promising object detection performance, but it still faces challenges in cap-

收稿日期:2025-09-10;**修回日期**:2025-11-20

*通信作者:欧阳晖 472321742@stu.lntu.edu.cn

基金项目:国家自然科学基金(61172144)

Supported by:National Natural Science Foundation of China(61172144)

turing fine-grained details of objects at multiple scales, particularly for small objects, and in effectively leveraging frequency-domain information. To address these limitations, we propose an enhanced RT-DETR architecture that integrates several innovative modules and techniques. For improved backbone efficiency and feature propagation, our model incorporates Cross Stage Partial (CSP) connections, which strengthen gradient flow across network stages and reduce redundant computations without sacrificing representation power. Building on this foundation, we introduce a Frequency Aware Module (FAM) that explicitly harnesses high- and low-frequency visual cues to enrich the feature representations, ensuring that important textural and edge details are preserved and emphasized throughout the network. At the feature interaction level, we design a Frequency-Enhanced Attention-based Intra-scale Feature Interaction (FAIFI) mechanism. By enabling features within the same scale to communicate through attention operations modulated by frequency-domain signals, FAIFI allows the detector to better capture subtle object characteristics that might otherwise be missed in single-scale processing. Complementing this, we employ a Boundary-Aware Feature Aggregation (BAFA) strategy to improve multi-scale feature fusion. BAFA specifically focuses on preserving object boundary information when combining features from different resolution levels, which helps maintain precise localization cues and contributes to more accurate detection of objects with distinct edges. In addition to these architectural improvements, we refine the training process with an Inner-GIoU loss, a modified bounding box regression loss that places greater emphasis on the overlapping area between predicted boxes and ground-truth boxes. This tailored loss function drives the model to achieve tighter and more reliable localization. Furthermore, to bolster the model's robustness and expand the diversity of training data, we introduce a novel frequency-based data augmentation approach utilizing the two-dimensional Discrete Wavelet Transform (2D-DWT). By decomposing and altering the frequency components of training images, this augmentation method exposes the network to a broader variety of spatial-frequency patterns and helps prevent overfitting to common textures. Together, these enhancements significantly boost the detection accuracy and reliability of RT-DETR. Experiments on challenging object detection benchmarks indicate that the improved RT-DETR not only achieves higher average precision than its baseline counterpart, especially on small or texture-rich objects, but also maintains real-time inference speed. Our results underscore the benefit of incorporating frequency-domain insights and boundary-aware feature processing into the RT-DETR framework for high-performance object detection. **Result** We rigorously evaluated our approach on three widely used person-detection benchmarks—CityPersons, WiderPersons, and ACDC. On CityPersons, our FEBA-DETR model achieved notable improvements over the baseline RT-DETR, with AP50 increasing by 2.3 percentage points and the stricter AP50:95 by 2.0 points. Furthermore, incorporating our data augmentation technique boosted performance, raising these gains to 3 and 2.4 points, respectively. Notably, on the occlusion-specific subsets of CityPersons, the miss rate (MR^{-2}) dropped by 4.92% for the partially occluded "Reasonable" category and by 2.82% for the heavily occluded "Heavy" category. Similarly, on the challenging WiderPersons dataset, FEBA-DETR again surpassed the baseline, with AP50 and AP50:95 improvements of 1.1% and 0.7%, respectively. These consistent gains underscore the robust generalization of our method across diverse datasets. Additionally, on the ACDC dataset, which specifically assesses robustness in complex environmental scenarios, FEBA-DETR boosted AP50 and AP50:95 by 1.9% and 1.3%, respectively, with even greater improvements of 3.1% and 2.4% after applying our data augmentation strategy. Collectively, these extensive experimental results convincingly affirm the effectiveness of our multi-level frequency enhancement and boundary-awareness techniques, significantly advancing pedestrian detection in challenging autonomous driving contexts. **Conclusion** Experimental evaluations show that FEBA-DETR delivers notable advancements in detecting pedestrians, particularly those who are either small in scale or partially occluded. The model also maintains consistent performance under more challenging and dynamic conditions. These findings underscore the practical advantages of incorporating frequency-domain enhancements along with boundary-awareness strategies. Together, these design choices contribute to a more dependable and accurate pedestrian detection system, which is especially valuable for autonomous driving applications where safety and precision are critical.

Key words: autonomous driving; pedestrian detection; small target detection; frequency domain enhancement; boundary awareness; data augmentation

0 引 言

行人检测是计算机视觉的一项关键任务，准确可靠的行人检测算法在自动驾驶、智能监控等领域起着至关重要的作用。尤其在自动驾驶领域，自动驾驶系统需要传感器实时捕获并精准处理复杂多变的交通场景信息，对行人检测算法的准确性、鲁棒性与实时性有更高的要求。

早期的行人检测方法主要依赖于手工设计特征，有研究者（Dalal N等，2005）利用方向梯度直方图（histogram of oriented gradient，HOG）来进行行人检测，接着有研究者（Wang X等，2009）在此基础上结合局部二值模式改进HOG方法，提出了一种能够检测局部遮挡的行人检测模型。还有研究者（Dollár P等，2009）利用积分通道特征检测方法，通过对图像通道积分特征的提取与分析，来提高行人检测的准确性和效率。然而，上述手工设计特征的方法精度低且不具备泛化性。

近年来，深度学习技术的突破为行人检测提供了更为精准和高效的解决方案。许多研究者开始将基于CNN的深度学习方法如Faster R-CNN（Ren等，2016）、YOLO（Redmon等，2016）和SSD（Liu等，2016）等改进或直接应用到行人检测领域。有研究者（Wei R等，2020）通过在主干网络中融合浅层和深层特征并引入多尺度融合策略，从而降低对小目标和遮挡目标的误检率。还有研究者（Ghari B等，2024）利用可见光与红外多光谱信息的融合，以提高弱光或恶劣天气条件下的行人检测性能。通过对模型的不断改进，基于CNN的检测模型在行人检测领域取得了不错的成效，但这种方法仍存在一些局限：卷积操作固有的局部感受野限制（Luo等，2016）使模型难以捕获图像的全局上下文信息，导致在处理尺度变化剧烈和被严重遮挡的行人目标时性能下降。

为解决CNN模型感受野受限的问题，近年来Transformer架构被引入目标检测任务。基于Transformer的目标检测模型，如端到端的目标检测算法DETR（Detection Transformer）在提高小尺度行人的检测精度和应对复杂遮挡场景方面展现出一定优势（Cao等，2022）。然而，Transformer类检测模型参数量大、计算复杂度高且训练收敛较慢，无法满足自动驾驶系统的高实时性要求。近期有研究致力于提升Transformer检测模型的推理效率，其中实时目标检测Transformer（Real-Time Detection Transformer，RT-DETR）作为一种端到端的实时Transformer检测算法被提出（Zhao Y等，2024）。RT-DETR采用了轻量化的CNN作为主干网络，用于提取图像特征，并将提取的特征输入到Transformer编码器-解码器结构中。编码器利用自注意力机制处理这些特征，接着解码器生成目标的边界框和类别预测。最终检测头再直接回归目标的边界框和类别。RT-DETR通过这种结构优化和轻量化设计大幅降低计算量，并在保持精度的同时将检测速度提升到与单阶段CNN检测器相当的水平。尽管RT-DETR在推理速度和效率方面表现出色，但该架构仍存在一些局限性。

RT-DETR在面对密集分布的目标、严重遮挡以及尺度剧烈变化时，仍然存在检测精度下降的问题。复杂环境干扰对模型性能的影响也未得到有效缓解。且现有的RT-DETR改进方法（Kong Y等，2024）也难以有效应对这些特殊挑战。而最近，频域特征的利用在目标检测领域展现出独特优势。有研究者（Wang等2023）在无人机检测中通过频域解缠模块将频域特征分离，以减小不同场景下的域偏差，增强了模型对小目标的感知能力。还有研究者（Zhong等，2022）提出在DCT频域中引入可学习频率增强模块用于挖掘目标在频谱域的隐藏线索，从而改进对复杂背景下目标的识别能力。然而，现有的方法未能充分挖掘频域与空间域在多层次特征交互中的协同潜力。

为此，本文针对行人检测任务对RT-DETR架构进行改进，提出面向自动驾驶复杂场景的实时Transformer行人检测算法FEBA-DETR，在保留Transformer全局建模优势的同时，针对小尺度行人和被遮挡行人进行改进，实现高效精确的检测。实验结果表明本文方法在提升检测精度和鲁棒性方面得到了一定的改善，验证了本文方法在自动驾驶场景行人检测任务中的有效性。

本文的主要贡献包括：提出频域感知模块（FAM），实现频域全局特征与空间域局部特征融合，提升远距离小目标的检测能力和抗噪能力；构建频域增强注意力特征交互模块（FAIFI），进一步强化频域与空间域特征的深层关联；提出边界感知特征重聚合机制（BAFA），着重建模多尺度特征图中的边界

细节，增强严重遮挡与背景干扰场景下行人目标的定位准确性；提出一种基于二维离散小波变换的频域子带数据增强方法，以应对自动驾驶中雨、雪、雾以及低光照等复杂场景。

1 FEBA-DETR算法

FEBA-DETR在RT-DETR基础上引入多重频域增强策略与边界感知机制，算法结构如图1所示。输入图像首先经过FAM-CSP骨干网络进行多尺度特征提取，并在FAM模块内部实现频域特征与空间特征的初步融合。随后，将FAM-CSP输出的多尺度特征图（P3、P4和P5）送入混合编码器（Hybrid Encoder）；在混合编码器中，仅将深层特征层P5输入FAIFI模块进行位置编码，并进一步强化频域特征与空间特征的深层交互。接着，再把处理后的P5特征和其余尺度的特征图一并送入BAFA特征聚合机制，在进行特征融合的同时突出目标边界细节，再将融合后的特征送入解码器。在最后的解码器阶段，采用RT-DETR中的不确定性最小化查询选择机制初始化解码器的目标查询，并通过解码器与辅助预测头输出最终的类别和边界框结果。

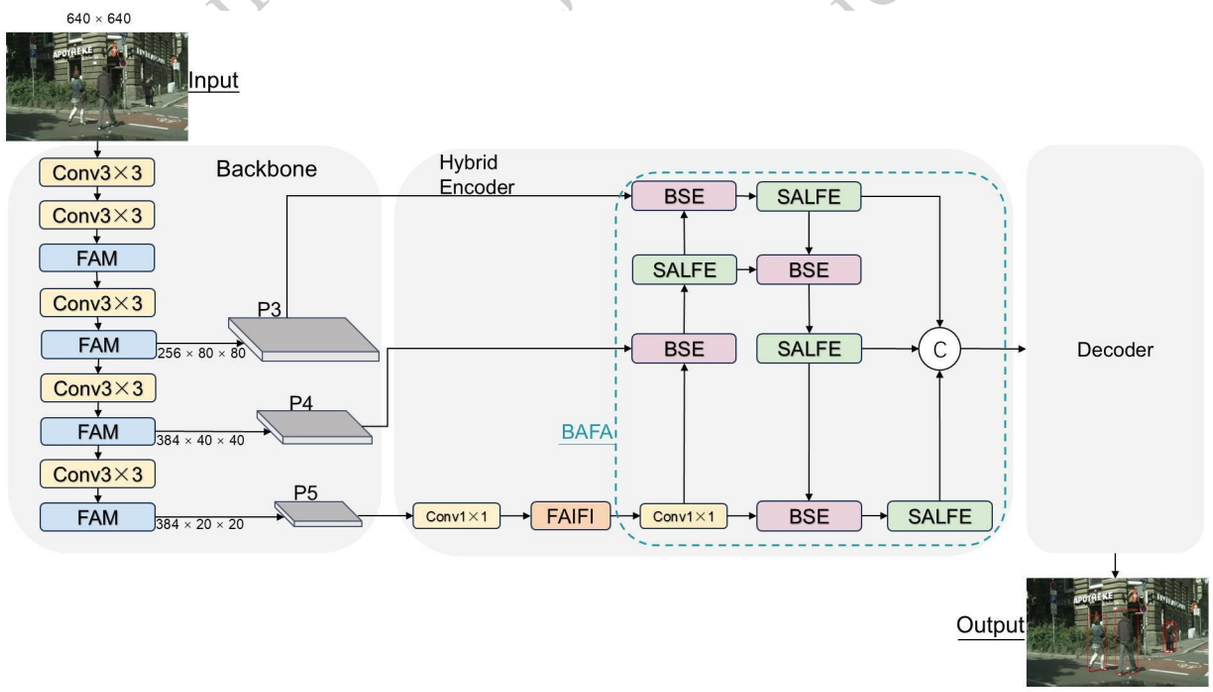


图1 FEBA-DETR结构图

Fig. 1 The Structure of FEBA-DETR

1.1 FAM-CSP骨干网络改进

高效的特征提取对行人检测任务至关重要（Khan等，2023），在复杂的自动驾驶场景中，行人目标在近距离和远距离的交通场景中尺度变化剧烈，且常常因噪声干扰而难以被有效识别。而传统卷积神经网络主要只依赖空间域的信息，往往忽略了跨尺度和全局信息（Carion N等，2020），由此可能导致在远距离小目标或者噪声干扰情况下检测效果不佳。原有RE-DETR结构在实时性和准确性取得了一定的平衡，但其Backbone同样也只侧重于空间域特征的融合。

为应对这一问题，本文在Backbone中提出频域感知模块，同时引入CSP的设计思想（Wang C Y等，2020）来进行改进。与传统主要侧重于空间域信息的卷积模块不同，FAM采用频-空域融合的思想。在FAM中，频域全局特征的引入不仅为模型提供了额外的上下文信息，提升了模型对目标尺度变化剧烈、细节信息匮乏的小目标行人感知能力。频域特征对图像退化成分（如噪声、模糊）固有的分离特性以及低频分量对局部干扰的抑制效应，还能在干扰较强的复杂环境中提供更加稳定和精准的特征表示。而CSP思想的引入使得Backbone在网络各个阶段实现更高效的信息传递与特征整合。

FAM分为三个部分：空间域特征提取部分、频域特征处理部分、跨域融合部分，具体结构如图2所示。

空间域特征提取部分对相同的两组输入特征 $\boldsymbol{F}_{in} \in \mathbf{R}^{B\times C\times H\times W}$ 采用两路并行的深度卷积来捕获不同感受野的空间特征，一路采用3×3深度卷积捕获细粒度局部特征，另一路采用5×5深度卷积增加感受野，以捕捉更多上下文信息。两路输出分别经过GeLU非线性激活函数增强特征表达能力，再通过

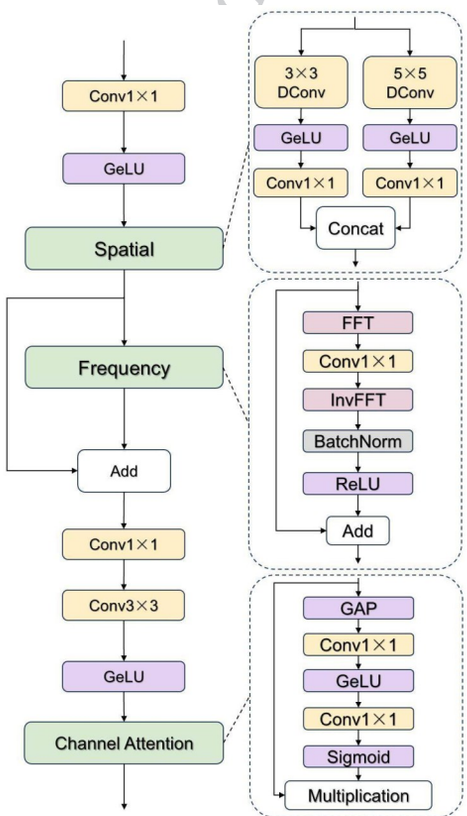


图2　FAM结构图

Fig. 2　The Structure of FAM

1×1卷积融合每个尺度的通道信息。最后在通道维度上进行拼接，以融合不同尺度的局部特征信息，最终得到空间域输出特征 $\boldsymbol{F}_{sp}$。公式表示为：

$$\boldsymbol{F}_{sp}=\mathrm{Concat}\begin{pmatrix}\mathrm{C}_{1\times1}\left(\mathrm{GeLU}\left(\mathrm{DC}_{3\times3}\left(\boldsymbol{F}_{in}\right)\right)\right),\\ \mathrm{C}_{1\times1}\left(\mathrm{GeLU}\left(\mathrm{DC}_{5\times5}\left(\boldsymbol{F}_{in}\right)\right)\right)\end{pmatrix} \tag{1}$$

式中，$\mathrm{C}_{1\times1}$ 表示1×1的卷积；$\mathrm{DC}_{k\times k}$ 表示深度可分离卷积操作，其中k为卷积核大小。

在FAM中，引入傅里叶变换来进行频域特征的处理主要出于两点优势考虑：一是它能够将图像中的退化成分有效分离。在真实道路环境中，干扰行人检测的常见因素通常会在频域中呈现出较为固定或易于识别的特征。通过在频域中对这些退化成分加以抑制，可以显著提高行人检测的准确率；二是傅里叶变换后的频域分量是由图像的所有空间分量共同参与计算得到的，这使得网络在在一定程度上拥有了类似自注意力机制的全局感受野，但它在保持全局建模能力的同时，计算开销却远低于自注意力机制（Lee Thorp J等，2021）。因此，在FAM的频域特征处理过程中，本文采用快速傅里叶变换（FFT）对前述输出的空间域特征进行转换来获得频域特征表示。具体来说，$\boldsymbol{F}_{sp}$ 经过快速傅里叶变换，获得频域特征表示。接着再经过1×1卷积操作进行频域上的特征整合处理，再通过逆傅里叶变换将特征还原到空间域。最后输出经过批归一化（BatchNorm）与ReLU激活函数，并通过残差连接将频域增强后的特征与原始空间特征相结合，最终得到频域特征 $\boldsymbol{F}_{freq}$：

$$\boldsymbol{F}_{freq}=\phi(\mathcal{F}^{-1}(\mathrm{C}_{1\times1}(\mathcal{F}(\boldsymbol{F}_{sp}))))+\boldsymbol{F}_{sp} \tag{2}$$

式中，$\mathcal{F}$ 与 $\mathcal{F}^{-1}$ 分别表示傅里叶变换与逆傅里叶变换操作；ϕ 表示BatchNorm与ReLU函数的组合。

跨域融合部分负责将处理后的频域特征与原始空间域特征进行结合。融合过程首先将空间域特征和频域特征进行相加，然后通过通道注意力模块生成自适应权重并对相加后的特征进行加权：

$$\begin{aligned}\alpha&=\sigma(\mathrm{C}_{1\times1}(\mathrm{GeLU}(\mathrm{C}_{1\times1}(\mathrm{GAP}(\boldsymbol{F}_{sp}+\boldsymbol{F}_{freq})))))\\ \boldsymbol{F}_{out}&=\alpha\odot(\boldsymbol{F}_{sp}+\boldsymbol{F}_{freq})\end{aligned} \tag{3}$$

式中，GAP表示全局平均池化；σ表示Sigmoid激活函数；⊙表示逐元素乘法；α表示通道注意力权重，$\alpha\in\mathbf{R}^{B\times C\times1\times1}$。

经此过程，FAM模块的输出特征既有空间域域局部细节，又包含频域全局纹理信息。从而有效提升Backbone对远距离小目标的检测能力和抗噪能力。

1.2　FAIFI模块

RT-DETR的AIFI模块（Attention-based Intra-scale Feature Interaction）仅在最高特征层进行自注意力计算。这种设计虽然大幅降低了计算开销，但由于忽视了中低层特征的注意力建模，使得细节信息的捕获主要依赖后续卷积跨尺度融合（CCFF）模块完成。这种策略在远距离小尺度行人检测任务中存在局限性，因为小目标往往仅在中低层特征层中具备显著的表达，缺乏足够的自注意力机制支撑，可能导致特征表征不足，使得小目标的漏检或定位不

准的情况增加。

为了克服上述局限性，本文构建一种适合复杂自动驾驶场景的频域增强注意力特征交互模块（Frequency-aware Attention-based Intra-scale Feature Interaction，FAIFI），从空间域、频域两个方面深化特征交互，提升特征表示的丰富性和判别力。引入频率自注意力（Frequency Self-Attention，FSA）（Zhang F等，2023）和空间自注意力（Spatial Self-Attention，SSA）（Alfasly S等，2024）两个分支，这两个分支的输出特征再由本文提出的频空特征融合网络（Frequency-Spatial Fusion Network，FSFN）进行融合。FSA通过对特征图施加快速傅里叶变换，以捕捉不同频段信号之间的关系，使得模型能够关注行人目标与背景在纹理细节上的差异而非仅语义类别。SSA则类似于原有的AIFI，用以捕获空间范围内的长距离依赖关系和局部细节。最后FSFN将这两种不同域的信息高效融合，以进一步深化频域与空间域之间的高效特征交互。

1.2.1　FSA

FSA首先通过快速傅里叶变换将输入特征$\boldsymbol{X}\in\mathbf{R}^{B\times C\times H\times W}$转换为频域查询特征$\boldsymbol{Q}_f$、键特征$\boldsymbol{K}_f$和值特征$\boldsymbol{V}_f$。接着将查询特征$\boldsymbol{Q}_f$和键特征$\boldsymbol{K}_f$进行注意力计算得到注意力图$\boldsymbol{A}_f$。再将其拆分成实部和虚部分别进行Softmax归一化后再重新组合，得到归一化后的频域注意力矩阵$\boldsymbol{Y}_f$。最后通过加权操作将其与值特征$\boldsymbol{F}_V$结合后再通过逆傅里叶变换和模运算得到最终的频域注意力特征$\boldsymbol{Y}$。公式表示为：

$$\begin{aligned}\boldsymbol{Y}_f&=\mathrm{complex}\left(\mathrm{Softmax}\left(\mathrm{Re}\left(\boldsymbol{A}_f\right)\right)+\boldsymbol{i}\mathrm{Softmax}\left(\mathrm{Im}\left(\boldsymbol{A}_f\right)\right)\right)\\\boldsymbol{Y}&=\left|\mathrm{IFFT}\left(\boldsymbol{Y}_f\cdot\boldsymbol{F}_V\right)\right|\end{aligned}\tag{4}$$

式中，Re与Im分别表示复数矩阵的实部与虚部；$\boldsymbol{i}$表示虚部单位；complex表示实虚矩阵的复数组合操作；$|\cdot|$表示模运算。

同时，引入残差通路，利用Sigmoid函数实现门控机制，进一步强调频域信息。最终进行维度拼接和1×1卷积融合后，获得频域增强的输出特征$\boldsymbol{F}_{out}$。公式表示为：

$$\begin{aligned}\boldsymbol{M}&=\sigma(W_2(\varphi(W_1(\mathcal{F}(\boldsymbol{F}_{in})))))\\\boldsymbol{R}&=\left|\mathcal{F}^{-1}(\boldsymbol{M}\odot\mathcal{F}(\boldsymbol{F}_{in}))\right|\\\boldsymbol{F}_{out}&=\mathrm{C}_1(\mathrm{Concat}(\boldsymbol{Y},\boldsymbol{R}))\end{aligned}\tag{5}$$

式中，$\boldsymbol{M}$表示经过两次1×1卷积、BN、ReLU以及Sigmoid生成的通道门控图；φ表示批归一化和ReLU函数的组合操作；$\boldsymbol{R}$表示残差通路的输出结果；W_1和W_2表示可学习的1×1卷积。

1.2.2　SSA

SSA首先对输入特征$\boldsymbol{X}$采用两种不同感受野的深度可分离卷积（分别为3×3和5×5卷积）提取多尺度的查询和键特征。随后将查询$\boldsymbol{Q}_S$和键$\boldsymbol{K}_S$特征进行维度重塑，计算它们之间的空间注意力矩阵，并对其应用Softmax函数归一化，得到空间自注意力权重。公式表示为：

$$\boldsymbol{X}_S^{att}=\mathrm{Softmax}\left(\boldsymbol{Q}_S\boldsymbol{K}_S^{\mathrm{T}}\right)\boldsymbol{V}_S\tag{6}$$

利用该空间注意力权重对值特征进行加权融合，形成空间注意力增强的特征图。然后再通过残差连接和不同卷积核提取特征，保留更多局部信息。最后通过拼接与1×1卷积融合得到最终的空间域增强输出特征。公式表示为：

$$\begin{aligned}\boldsymbol{P}_{res}&=\mathrm{Concat}(\mathrm{DC}_{3\times3}(\boldsymbol{F}_{in}),\mathrm{DC}_{5\times5}(\boldsymbol{F}_{in}))\\\boldsymbol{F}_{out}&=\mathrm{C}_{1\times1}\cdot\mathrm{Concat}(\boldsymbol{X}_s^{\,att},\boldsymbol{P}_{res})\end{aligned}\tag{7}$$

式中，$\mathrm{DC}_{k\times k}$表示深度可分离卷积操作，其中k为卷积核大小。

1.2.3　FSFN

FSFN结构如图3所示。首先将FSFN的输入特征$\boldsymbol{F}_{in}$进行层归一化，得到归一化特征$\boldsymbol{X}$。再将$\boldsymbol{X}$送入不同的处理分支进行高效融合。

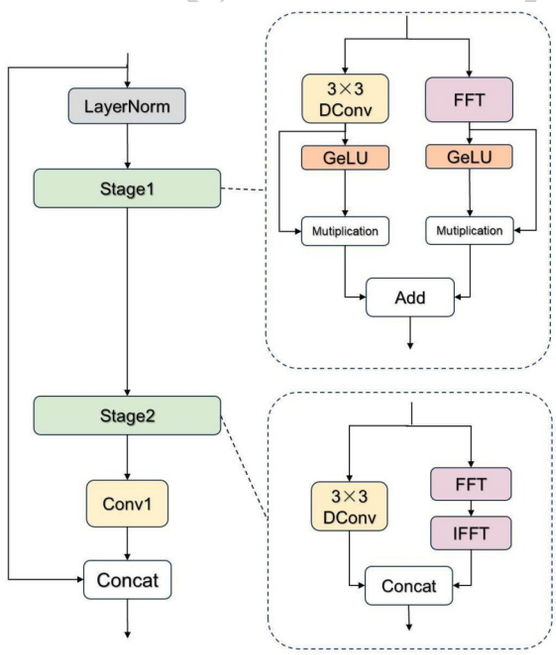


图3　FSFN结构图

Fig. 3　The Structure of FSFN

在第一阶段的频域处理分支中，首先通过FFT操作将归一化特征$\boldsymbol{X}$转换到频域，然后通过GeLU激活产生门控信号与频谱逐元素相乘，以突出具有判别力的频率成分，从而得到频域中间特征$\boldsymbol{X}_f$；在空间域分支，使用深度可分离卷积提取空间局部结构细节，再经过GeLU非线性激活进行自门控，以增强目标区域特征并抑制背景干扰，从而得到空间域中间特征$\boldsymbol{X}_s$。公式表示为：

$$\begin{aligned}\boldsymbol{X}_f &= \mathrm{GeLU}(\mathcal{F}(\boldsymbol{X}))\odot\mathcal{F}(\boldsymbol{X})\\ \boldsymbol{X}_s &= \mathrm{GeLU}(\mathrm{DC}_{3\times3}(\boldsymbol{X}))\odot\mathrm{DC}_{3\times3}(\boldsymbol{X})\end{aligned}\tag{8}$$

式中，$\mathrm{DC}_{k\times k}$表示深度可分离卷积操作，其中k为卷积核大小；$\mathcal{F}(X)$表示傅里叶变换操作；$\odot$表示逐元素乘法。

在获得中间特征$\boldsymbol{X}_f$和$\boldsymbol{X}_s$后，对其逐元素相加进行空间域与频域的初步融合，随后再进一步重构：一方面对该融合特征执行傅里叶变换，再通过逆傅里叶变换还原到空间域，得到频域重构特征$\boldsymbol{X}_f'$；另一方面，对该融合特征再施加一次深度可分离卷积得到空间域重构特征$\boldsymbol{X}_s'$。公式表示为：

$$\begin{aligned}\boldsymbol{X}_f' &= \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{X}_s+\boldsymbol{X}_f))\\ \boldsymbol{X}_s' &= \mathrm{DC}_{3\times3}(\boldsymbol{X}_s+\boldsymbol{X}_f)\end{aligned}\tag{9}$$

在计算效率方面，FAIFI引入了双路径注意力，不可避免地增加了一些计算量。但是本文将频域通道维度降低到128，限制频域注意力的开销，并通过并行的、轻量化的FFT实现来减小影响。优化后的FAIFI模块的计算量基本与AIFI的计算量处于同一量级。

1.3 BAFA特征聚合机制

在自动驾驶场景中，行人目标往往频繁遮挡和受背景干扰，对此，精确定位行人目标的边界位置至关重要。然而，一般的跨尺度融合机制通常忽略了特征之间的语义差异与边界细节信息，导致模型在处理行人频繁遮挡和背景干扰较强的情况下出现较多的漏检和误检。

RT-DETR中的CCFF特征融合机制通过上采样、卷积和特征拼接，有效整合了来自不同尺度的特征，并保留了全局语义信息，在一定程度上缓解了上述问题。然而，CCFF对行人局部特征捕捉不够精准，在边界细节信息的提取上不够精细。尤其是在遮挡和背景干扰较强的情况下，对行人特征的捕捉仍存在遗漏。

本文提出一种边界感知特征聚合机制（Boundary-Aware Feature Aggregation，BAFA），以解决CCFF在检测中边界细节丢失和特征融合不足的问题。BAFA由选择性边界增强模块（Boundary-Selective Enhancement，BSE）与移位感知提取模块（Shift-Aware Extraction，SAE）构成。

1.3.1 BSE模块

BSE模块通过引入SBA（Tang F等，2023）的自适应边界聚合思想（Selective Boundary Ag gregation），自适应融合浅层边界信息与深层次语义特征，有效增强目标的边界定位能力。

整个模块由两个并行的重校准注意力单元（Recalibration Attention Unit，RAU）组成，两个并行的RAU单元分别处理来自边界特征图$\boldsymbol{F}_b$和语义特征图$\boldsymbol{F}_s$的互补特征。一个是以边界特征为主、语义特征为辅的$\boldsymbol{R}_{high}$，一个是以语义特征为主、边界特征为辅的$\boldsymbol{R}_{low}$，再通过通道维度拼接，并以3×3卷积精炼聚合，产生最终输出特征图：

$$\begin{aligned}\boldsymbol{R}_{high} &= \mathrm{RAU}(\boldsymbol{F}_b,\boldsymbol{F}_s)\\ \boldsymbol{R}_{low} &= \mathrm{RAU}(\boldsymbol{F}_s,\boldsymbol{F}_b)\\ \boldsymbol{Z} &= \mathrm{C}_{3\times3}\left(\mathrm{Concat}\left(\boldsymbol{R}_{high},\boldsymbol{R}_{low}\right)\right)\end{aligned}\tag{10}$$

式中，RAU表示RAU单元。$\mathrm{C}_{3\times3}$表示3×3的卷积。

其中RAU结构如图4所示。RAU接受一对输入特征图$\boldsymbol{T}_1$和$\boldsymbol{T}_2$，首先通过1×1的卷积映射并经过sigmoid激活，生成对应的注意力权重映射$\boldsymbol{T}_1'$和$\boldsymbol{T}_2'$，如下式所示：

$$\begin{aligned}\boldsymbol{T}_1' &= \sigma\left(W_1*\boldsymbol{T}_1\right)\\ \boldsymbol{T}_2' &= \sigma\left(W_2*\boldsymbol{T}_2\right)\end{aligned}\tag{11}$$

式中，W_1、W_2为可学习的1×1的卷积操作。

接着计算这两个注意力图的差异，并且为进一步强化差异，加入反转差异特征Reverse分支：

$$\begin{aligned}\boldsymbol{D} &= \boldsymbol{T}_1' - \boldsymbol{T}_2'\\ \boldsymbol{T}_2'' &= 1 - \boldsymbol{T}_2'\end{aligned}\tag{12}$$

最终利用上述注意力和差异信息，对原始特征进行加权融合以突出边界区域的特征，得到RAU的输出特征：

$$\boldsymbol{R} = \left(\boldsymbol{T}_1'\odot\boldsymbol{T}_1\right) + \left(\boldsymbol{T}_1''\odot\boldsymbol{T}_2\right) + \boldsymbol{D}\tag{13}$$

式中，$\odot$表示逐元素相乘。

公式（15）表明，RAU一方面保留了由自身注意力加权的$\boldsymbol{T}_1$，另一方面融合了有反转差异权重$\boldsymbol{T}_1''$的$\boldsymbol{T}_2$特征，并结合了边界与语义注意力图之间的差异$\boldsymbol{D}$，这种融合策略使得边界显著处的差异得到强化，

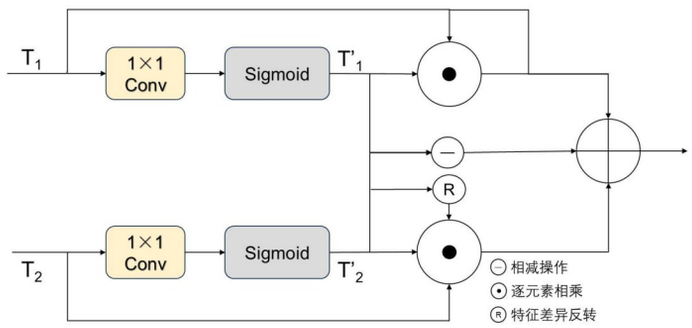


图4　RAU结构图

Fig. 4　The Structure of RAU

从而弥补了高层语义特征中缺失的边界细节，同时也为低层特征引入了高层语义校正。

1.3.2　SAE模块

SAE模块通过重复堆叠结构与移位卷积操作（Zhang X等，2018），进一步强化BSE模块提取的边界细节特征，使整体网络在边界感知方面更加精准，结构如图5所示。

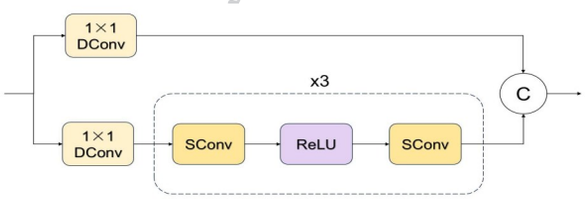


图5　SAE结构图

Fig. 5　The Structure of SAE

SAE模块首先将输入特征图分成两个分支，并分别通过1×1深度卷积，其中一路分支特征被送入由移位卷积（Shifted Convolution，SConv）与非线性激活函数（ReLU）组成的重复堆叠结构中，以显著扩大局部感受野，并反复强化对空间细节的表达能力。最后，再将该分支提取到的丰富空间细节特征与另一浅层分支特征直接拼接融合，公式表示为：

$$\begin{aligned} \boldsymbol{X}_1' &= \mathrm{DC}_{1\times1}(\boldsymbol{X}_1) \\ \boldsymbol{X}_2' &= \mathrm{DC}_{1\times1}(\boldsymbol{X}_2) \\ \boldsymbol{F} &= \mathrm{Concat}\left(\boldsymbol{X}_1', \mathrm{Stacked}_3 - \mathrm{SConv}\left(\boldsymbol{X}_2'\right)\right) \end{aligned} \tag{14}$$

式中，$\boldsymbol{X}_1$，$\boldsymbol{X}_2$为原始输入特征图经通道分割后的两个特征子集，$\mathrm{DConv}_{1\times1}$表示1×1深度卷积，用于对特征通道进行重构与轻量化变换；$\mathrm{Stacked}_3-\mathrm{SConv}\left(\boldsymbol{X}_2'\right)$表示对特征$\boldsymbol{X}_2$经过三次重复的移位卷积与非线性激活函数的组合操作。

通过这种设计，SAE模块增强模型对局部边缘轮廓与纹理细节的敏感性，尤其是在面对行人目标频繁遮挡和复杂背景干扰的情况时。与前述BSE模块紧密协同，提升整体模型的边界定位与检测性能。

1.4　频域子带数据增强

目前，大多数目标检测算法都会采用纯空间域的数据增强手段（Krizhevsky A等，2017），例如随机裁剪、水平翻转以及比例缩放等（Shorten C等，2019），RT-DETR也沿用了这些方法。这些空间域的增强方法优点是实现简单、计算高效，但在自动驾驶场景下进行行人检测仍存在不足，尤其是在雨、雾、雪等恶劣天气以及低光照环境下，难以有效弥补图像由外部环境干扰而造成细节特征退化，从而限制了检测模型的性能。

为解决这一问题，本文提出一种基于二维离散小波变换（2D-DWT）的频域子带数据增强方法，该方法核心思想是将图像在频域中分解为多个频率子带，并针对不同子带的系数进行有针对性的增强，从而强化图像的结构信息与细节纹理。具体来说，低频子带主要包含图像的整体亮度和对比度信息，通过对其进行增强，可以改善图像的全局视觉效果。而高频子带则包含图像的边缘和细节信息，增强高频子带有助于恢复和强化因环境干扰造成退化的图像细节特征。整个数据增强流程主要包括分解、增强和重构这三部分，接下来将分段介绍。

在分解部分，将输入RGB图像转换为灰度图$\boldsymbol{I}_{gray}$，并采用Haar小波进行一层二维离散小波分解，得到一个低频子带（包含整体结构与光照信息）$\boldsymbol{LL}$以及三个高频子带（反映图像细节与边缘信息）$\boldsymbol{LH}$、$\boldsymbol{HL}$和$\boldsymbol{HH}$。公式表示为：

$$[\boldsymbol{LL},(\boldsymbol{LH},\boldsymbol{HL},\boldsymbol{HH})] = \mathrm{haar}(\boldsymbol{I}_{gray}) \tag{15}$$

式中，haar表示二维Haar小波分解算子。

在增强部分，根据场景需求对各子带系数进行有针对性的增强处理，得到$\boldsymbol{LH}'$、$\boldsymbol{HL}'$、$\boldsymbol{HH}'$、$\boldsymbol{LL}'$。低光照场景下，采用直方图均衡函数CLAHE对低频子带$\boldsymbol{LL}$进行动态范围拉伸，以提升整体亮度和对比度；针对雨、雾、雪等能见度低的复杂场景，对高频子带系数实行放大操作，来增强受退化影响的细节纹理，并突出边缘信号来弥补轻量运动模糊造成的高频细节损失。公式表示为：

$$\begin{aligned}\boldsymbol{LH}' &= \alpha \cdot \boldsymbol{LH}\\ \boldsymbol{HL}' &= \alpha \cdot \boldsymbol{HL}\\ \boldsymbol{HH}' &= \alpha \cdot \boldsymbol{HH}\\ \boldsymbol{LL}' &= \beta \cdot \mathrm{CLAHE}(\boldsymbol{LL})\end{aligned} \tag{16}$$

式中，α为高频超参数；β为低频超参数。

在重构部分，通过逆小波变换重构增强后的灰度图$\boldsymbol{I}'_{gray}$，然后对所得灰度图的像素值进行裁剪，将所有像素限制在有效强度范围(0,255)以避免增强造成的溢出失真。最后将增强的灰度图转换回BGR彩色图像，得到最终增强图像$\boldsymbol{I}'$。公式表示为：

$$\begin{aligned}\boldsymbol{I}'_{gray} &= \mathrm{haar}(\boldsymbol{LH}',\boldsymbol{HL}',\boldsymbol{HH}',\boldsymbol{LL}')\\ \boldsymbol{I}'_{gray} &= \mathrm{clip}(\boldsymbol{I}'_{gray},0,255)\\ \boldsymbol{I}' &= \mathrm{GRAY2BGR}(\boldsymbol{I}'_{gray})\end{aligned} \tag{17}$$

经过此步骤处理后，图像的关键细节得到凸显，同时总体亮度和对比度均得到改善，可直接用于后续的行人检测算法输入。

2 实　验

2.1 数据集介绍

本文实验主要采用CityPersons数据集(Zhang S等，2017)、WiderPerson数据集(Zhang S等，2020)和ACDC (Adverse Conditions Dataset with Correspondences)数据集(Sakaridis C等，2021)来验证模型改进的有效性。

CityPersons数据集是从城市街景数据集Cityscapes中挑选出来的一个子集，涵盖3个国家、18个城市和3个季节，共包含5000张高分辨率图像(2048×1024)。其中训练、验证和测试集分别为2975张、500张和1575张，共含有行人目标35016个，平均每张图像包含7个行人目标。该数据集还额外提供了行人目标高度和可见度等数据。本文实验用该数据集作为基准数据集来评估算法改进的有效性。

WiderPerson数据集是当前规模最大的拥挤场景行人检测数据集之一，包含13,382张图像，共计约400000个不同遮挡的行人实例。其中训练集8000张、验证集1000张、测试集4,382张。该数据集包含行人、骑行的人、部分可见的人、人群和忽略区域5类标注，本文只采用前三类标注，并将行人、骑行的人和部分可见的人都归为一类。鉴于自动驾驶场景下的行人检测需要具备在真实复杂环境中的泛化能力，所以本文使用包含多种场景的WiderPerson数据集来验证改进后算法的泛化能力。

ACDC (Adverse Conditions Dataset with Correspondences)数据集专为在不利视觉条件下训练和测试语义感知方法而设计，涉及雾、雨、雪、夜视4类不利条件的多模态数据，包含4006张图像，每类场景提供1,500组同步RGB-热成像对齐数据。数据集真实模拟了传感器在恶劣环境中的性能退化效应。本文实验用该数据集来评估算法在恶劣天气和低光照条件下的行人检测能力。

2.2 实验环境与设置

本文实验环境如下：操作系统为Ubuntu 20.04操作系统，处理器为Intel Xeon Platinum8362，内存配置为32GB，深度学习框架选用PyTorch-1.12.1，图形处理器为24GB显存的NVIDIA GeForce RTX 3090 GPU，cuda版本为11.3。在实验设置方面，输入图像设置为640×640，训练轮数为350轮，初始学习率为0.0002，批量大小为8，优化器为AdamW，其余设置均与原RT-DETR算法相同。

2.3 实验评估指标

本文实验主要采用平均精度(Average Precision，AP)和对数平均漏检率(Log-Average Miss Rate，MR^{-2})作为模型性能评估的核心指标。其中，前者衡量模型在多尺度目标上的检测准确性，后者评估模型在不同遮挡等级下的漏检情况。

AP通过计算精确度-召回率曲线下的面积来实现：

$$AP = \int_0^1 p(r)\,dr \tag{18}$$

式中，$p(r)$为在召回率r下对应的精确度。

本文主要关注两个AP指标，AP50与AP50:95。AP50对应当预测框与真实框之间的交并比阈值设为0.50时模型的平均精度表现；AP50:95对应当计

算IoU阈值在0.50到0.95(间隔为0.05,共10个阈值)下的平均精度的均值。

MR^{-2}是衡量目标检测模型在特定误检率范围内综合性能的指标,其核心在于量化漏检率(Miss Rate, MR)与单位图像误检率(False Positive Per Image, FPPI)之间的对数关系。MR^{-2}的计算流程如下:在以FPPI为横轴、log(MR)为纵轴绘制的曲线图上,在FPPI的10^{-2}至10^{0}的对数区间内均匀采样9个点,计算对应的漏检率并求平均,再对其指数变换最终得到MR^{-2}。其数学表达式为:

$$MR^{-2} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(MR\left(FPPI_i\right)\right)\right) \quad (19)$$

式中,n为9。

本文中MR^{-2}指标的评估遵循CityPersons数据集相关标准(Zhang S等,2017),将行人根据其遮挡情况分为下面三类子集:高度大于50像素且可视率大于65%的行人被划分为Reasonable子集;高度介于50-75像素且可视率大于65%的行人被划分为Reasonable_small子集;高度大于50像素且可视率介20%到65%的行人被划分为Reasonable_occ子集。本文实验只关注Reasonable与Reasonable_occ(为便于区分,后续实验记作Heavy)两类子集,以验证模型在轻微遮挡与严重遮挡场景下的检测性能。

2.4 对比实验

2.4.1 不同算法对比

为了验证FEBA-DETR的有效性,本节在CityPersons数据集上将其与两大类典型的实时目标检测算法进行系统对比。一类是基于CNN的实时检测模型YOLO系列,包括YOLOv5m、YOLOv8m、YOLOv10m、YOLOv11m和YOLOv12m,另一类是近年来兴起的基于Transformer架构的实时检测模型,包括当前备受关注且目前表现最为优异的D-FINE(Peng Y等,2024)、DEIM(Huang S等,2024)。为了实验对比的公平性,选取的这些对比模型的参数量和计算复杂度均与本文提出算法非常接近。

从表1所示的对比结果可以看出,本文所提出的方法在多个性能指标上都展现出了优势。在AP50和AP50:95指标上较原模型分别提高了2.3个百分点和2个百分点,并且超出了其他所有对比模型。在一般遮挡的Reasonable指标中,本文方法为21.73%,相对优于其他对比模型,较原模型降低了约5个百分点,说明本文方法对一般遮挡的行人目标检测具有更强的鲁棒性;而在严重遮挡的Heavy指标中,本文方法为50.82%,同样优于现有模型。同样的,表明本文方法在处理严重遮挡的行人目标时具有更低的漏检率。

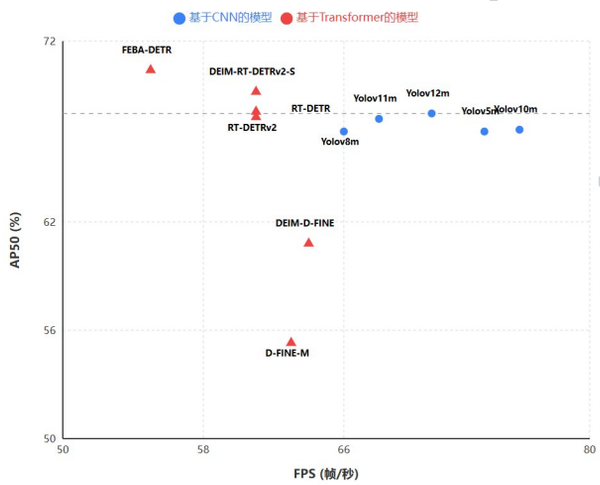


图6 精度速度对比图

Fig. 6 Accuracy-Speed Comparison

图6进一步展示了不同模型在精度与速度之间的权衡关系,从散点分布可以观察到,基于CNN的模型(蓝色圆点)主要集中在高速度区域(66-76 FPS),但AP50相对较低(67.0-68.0%);而基于Transformer的模型(红色三角)则正好相反。本文提出的FEBA-DETR位于图6的左上角,尽管在实

时性的FPS指标方面,本文方法达到55帧/秒,略低于其他模型,但仍然满足实时检测需求。更重要的是,在AP50指标上每提升1%的精度代价仅为约7 FPS的速度损失,这一精度-速度比显著优于现有方法,表明本文方法在保持较高检测精度的同时,实现了更优的综合性能平衡。

值得注意的是,同为Transformer检测模型的DEIM-D-FINE和D-FINE-M的检测精度远低于其余模型。经过多次实验和分析,这两个模型均采用HGNetv2作为主干网络,而其余性能较优的模型则使用了不同的backbone。实验结果表明,HGNetv2在本文的行人检测任务中表现欠佳,可能需要针对行人检测场景的特点进行专门的优化或选择更适配的主干网络。

为了体现结构改进的有效性,对比实验并未使用前面提到的频域子带数据增强方法,频域子带数据增强的改进有效性将在后续实验部分证明。

表1　不同算法对比表

Table 1　The Comparison Table of Different Algorithms

Model	Type	Params	GFLOPs(G)	AP50 (%)	AP50:95 (%)	Reasonable(%)	Heavy (%)	$FPS_{bs=1}$
Yolov5m	CNN	25M	64	67.0	43.6	29.51	60.01	74
Yolov8m	CNN	25M	78	67.0	43.5	28.71	59.02	66
Yolov10m	CNN	16M	63	67.1	43.5	31.84	59.66	**76**
Yolov11m	CNN	20M	67	67.7	44.8	28.68	55.88	68
Yolov12m	CNN	19M	60	68.0	44.7	27.33	56.46	71
D-FINE-M	Transformer	19M	56	55.3	30.1	25.52	52.84	63
DEIM-RT-DETRv2-S	Transformer	20M	60	69.2	43.5	23.72	52.13	61
DEIM-D-FINE	Transformer	19M	57	60.8	36.0	25.37	54.23	64
RT-DETR	Transformer	19M	57	68.1	42.2	26.65	53.64	61
RT-DETRv2	Transformer	19M	57	67.8	42.1	26.90	53.55	61
FEBA-DETR	Transformer	18M	63	**70.4**	**44.2**	**21.73**	**50.82**	55

注：加粗字体为每列最优值；AP50和AP50:95指标数值越高越优；Reasonable和Heavy数值越低越优。

2.4.2　不同数据集对比

本节在不同的数据集上进行了对比验证，结果如表2所示。从结果可以看出，本文提出的方法除了在基准数据集CityPersons上的AP指标分别提升了2.3%和2%，表明本文方法在复杂的城市交通场景中在处理小目标行人和遮挡问题方面表现出了更优异的性能。而在不仅限于交通场景的WiderPerson数据集上，其AP指标同样也超过了原模型1.1%与0.7%，充分展现了模型的泛化能力。在ACDC数据集上分别提升了1.9%和1.3%，证明了算法在复杂环境下尤其是在不利天气以及低光照条件下的行人检测性能。

2.4.3　不同损失函数对比

为了探究不同的边界框回归损失函数在行人检测任务中的性能差异，本节在相同实验设置下，对多种主流回归损失函数在CityPersons数据集上进行横向对比，结果如表3所示。

表2　不同数据集对比表

Table 2　The Comparison Table of Different Datasets

Datasets	Model	AP50(%)	AP50:95(%)
CityPersons	RT-DETR	68.1	42.2
	FEBA-DETR	70.4	44.2
WiderPerson	RT-DETR	78.8	49.2
	FEBA-DETR	79.9	49.9
ACDC	RT-DETR	38.7	17.2
	FEBA-DETR	40.6	18.5

表3　不同损失函数对比表

Table 3　The Comparison Table of Different Loss

Loss	AP50 (%)	AP50:95(%)	Reasonable(%)	Heavy (%)
GIoU	70.0	43.9	22.62	**50.37**
MPDIoU	**70.4**	**44.3**	22.16	51.20
Focaler-GIoU	69.9	43.8	24.15	53.96
Wise-GIoU	69.3	43.8	23.90	52.58
Focaler-MPDIoU	69.6	43.7	23.12	52.77
Inner-MPDIoU	69.5	43.4	23.47	51.04
Inner-GIoU	**70.4**	44.2	**21.73**	50.82

注：加粗字体为每列最优值。

实验以RT-DETR算法原本采用的GIoU损失函数作为基准损失函数，比较的损失函数包括基于最小点距离度量来综合考虑重叠面积、中心点距离和宽高差异等几何因素的MPDIoU损失函数（Ma S等，2023）；引入了类Focal Loss的对焦机制以强化优化困难样本的Focaler-GIoU损失函数（Zhang H等，2024）；采用动态非单调的梯度权重分配策略，对不同质量的预测框进行“智慧”增益调节的Wise-GIoU

损失函数(Tong Z等,2023);以及体现了对边界框内部重叠结构的感知,通过辅助框机制自适应不同回归状态的Inner-GIoU损失函数(Zhang H等,2023)。此外,为进一步丰富比较范围,还有上述损失函数的组合改进,如将对焦机制与MPDIoU相结合的Focaler-MPDIoU损失函数(Zhang H等,2024)和将内部结构感知融入MPDIoU的Inner-MPDIoU损失函数(Ye R等,2024)。

从整体检测精度来看,MPDIoU及Inner-GIoU表现相对突出,特别是在高阈值场景下依然能保持较高检测精度,说明其在对边界框的准确度和可回归性上具有较好的平衡能力。再结合各项指标所体现的抗遮挡能力,MPDIoU尽管在轻微遮挡场景中表现出一定的优势,但在严重遮挡情形下明显不及Inner-GIoU。所以综合来看Inner-GIoU的提升更为显著,这表明其能够在严重遮挡或边界模糊的行人目标上施加更有效的约束,有助于模型在细粒度定位与边界识别阶段作出更精确且鲁棒的判断。

2.4.4 频域子带增强超参对比

本节通过展开对比实验来分析低频与高频超参数对频域子带数据增强数性能的影响,并在CityPersons与ACDC数据集上进行相应的评估,结果如表4所示。表中的α和β分别表示公式(16)提到的高频超参数和低频超参数。实验1对应在改进后的算法中,沿用原始RT-DETR的数据增强策略所得结果。实验2则对应使用频域子带数据增强方法但对核心超参不做额外调整的情形,即低频超参数和高频超参数均为1。而其余实验则是分别设置不同的低、高频系数所得到的结果。

表4 不同核心超参对比表

Table 4 The Comparison Table of Different Core Hyper-parameters

实验编号	超参数(α,β)	CityPersons		ACDC	
		AP50(%)	AP50:90(%)	AP50(%)	AP50:90(%)
1	–	70.4	44.2	40.6	18.5
2	α=1.0,β=1.0	70.8	44.4	41.3	19.0
3	α=1.2,β=1.0	70.8	44.5	41.6	19.2
4	α=1.0,β=1.1	71.0	44.6	41.8	19.4
5	**α=1.2,β=1.1**	**71.1**	**44.6**	**41.8**	**19.6**
6	α=1.5,β=1.2	69.5	43.1	39.1	17.8

由表4中实验2的数据可以看出,即使不对低频和高频系数进行额外的调节,本文的频域子带数据增强方法仍然比原有的数据增强提高了0.4%和0.2%,在ACDC数据集上提高了0.7%和0.5%。这一结果表明,即便在未进行精细调参的前提下,该方法也能够有效提升模型在一般场景和复杂场景下的检测表现。在实验3中,仅提升高频超参数的情况下,CityPersons数据集的AP50:90的表现略有提升,而在ACDC数据集上的指标进一步分别提升了0.3%和0.2%,这说明在以低光照或恶劣天气条件为主的复杂环境下,适度增强图像的全局亮度与对比度可更有效地改善模型的检测能力。同样的,在实验4中仅调节低频超参数,CityPersons的AP50与AP50:90的指标相较于实验2均提升了0.2%,而ACDC则分别提升了0.5%和0.4%,说明针对边缘信息的适度补偿对于复杂场景下的细节恢复可带来更有效的性能收益。在实验5中,同时调节低频和高频超参数后,CityPersons的AP50和AP50:90较实验2分别提升了0.3%和0.2%,较原始的数据增强方法(即实验1的数据)分别提升了0.7%和0.4%,ACDC数据集的提升则更明显,相较于实验2分别提升了0.5%和0.6%,较实验1分别提升了1.2%和1.1%。这些结果充分表明,高低频增强的联合策略具有协同效应,特别是在复杂场景下,二者的协同配合可显著提升图像可辨性和检测精度。这里实验5的增强策略也将作为本文后续实验的默认超参数配置,为后续实验提供统一的增强方案。

值得注意的是,当增强强度进一步增大时,即用实验6的这组超参数时,模型性能出现了下降。CityPersons的两个指标分别下降了0.9%和1.1%,ACDC数据集则下降了1.5%和0.7%。经过多次实验分析,过高的增强系数会引起过曝、噪点以及伪边缘等问题,导致图像原本的特征可靠性降低,进而影响最终的检测效果。此外,在实验过程中还发现,本文的频域子带数据增强方法还能在一定程度上小幅提升小目标与遮挡条件下的检测表现,分析认为,原因在于模型通过增强高频子带,能更好地提取到小尺度行人的关键结构信息,从而提高了部分被遮挡行人的识别能力。

2.5 消融实验

为了验证本文改进点的有效性,本节在CityPersons数据集上展开消融实验,如表5所示。从整体

结果来看，随着FAM-CSP、FAIFI、BAFA的逐步引入，算法在AP指标上均表现出稳定且显著的提升。其中，改进单一模块（实验2、3、4）在AP50最多提升1.1%，在AP50:95则最多提升1.2%，其中FAM的表现最为优异。而在MR^{-2}指标上表现最优异的是BAFA模块，其在严重遮挡的情况下Heavy指标下降了2.72%，在轻微遮挡指标Reasonable则下降了2.82%。双模块组合（实验5、实验6和实验7）进一步提高了检测精度，但在严重遮挡场景下仍存在小幅波动。

在引入全部改进项，并进一步结合替换损失函数与数据增强策略（即实验8至实验10）后，模型的整体检测性能达到了最优，AP50最高可达71.1%，同时在更严苛的衡量指标AP50:95上达到44.6%，相较于基线模型（实验1）提升幅度明显。此外，这种全模块设置在Reasonable指标中取得了最低的漏检率，仅为21.07%，表明模型对常见遮挡场景下的行人检测有更强鲁棒性。

值得注意的是，双模块组合在严重遮挡条件下的指标与三模块组合的差距并不明显，这也在一定程度上说明了特定模块在对复杂背景或部分遮挡场景的适用性有一定重叠。

表5　消融实验表

Table 5　The Table of Ablation Study

实验编号	FAM-CSP	FAIFI	BAFA	Loss	Aug	Params	GFLOPs (G)	Ap50 (%)	Ap50:95(%)	Reasonable(%)	Heavy (%)	$FPS_{bs=1}$
1						19873044	56.9	68.1	42.2	26.65	53.64	61
2	√					16559268	50.9	69.6	43.4	25.30	52.54	65
3		√				20118768	57.7	69.2	42.8	24.78	51.67	61
4			√			20581200	59.9	68.7	43.1	23.83	50.92	58
5		√	√			20725488	60.3	69.3	43.5	23.67	51.09	59
6	√		√			18611424	62.3	69.6	43.3	23.90	51.20	56
7	√	√				16804992	52.6	69.7	43.8	24.33	52.47	63
8	√	√	√			18340136	62.4	70.0	43.9	22.62	50.37	55
9	√	√	√	√		18340136	62.4	70.4	44.2	21.73	50.82	55
10	√	√	√	√	√	18340136	62.4	71.1	44.6	21.07	49.81	55

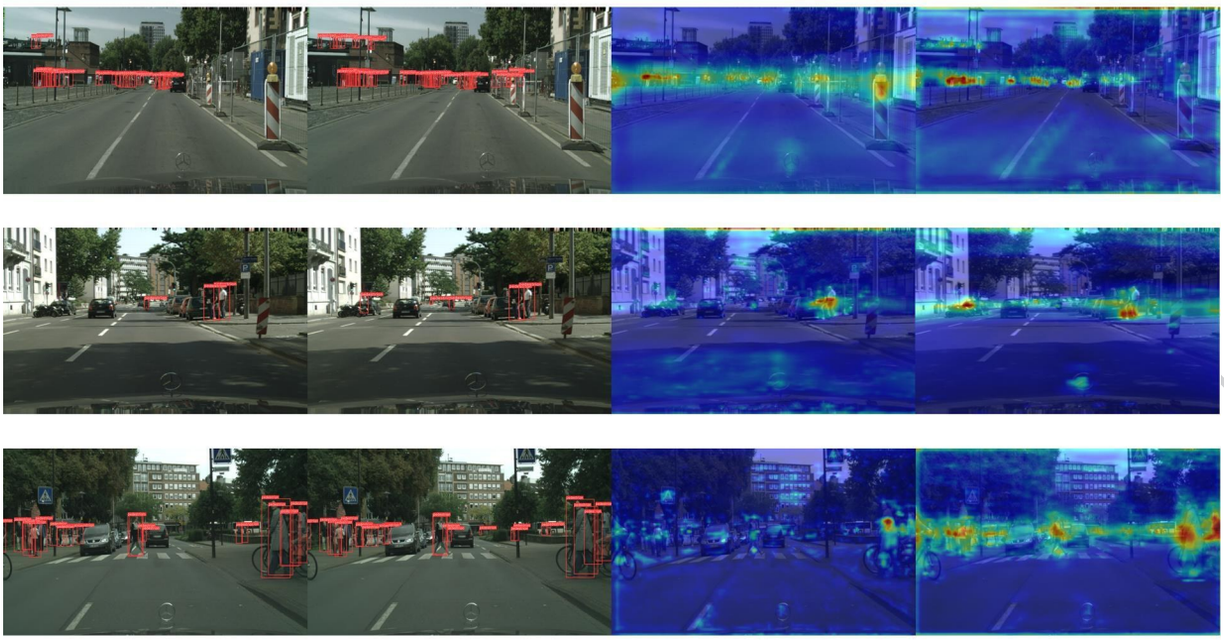
2.6　实验效果可视化

本文提出的改进算法与RT-DETR基准模型在CityPersons数据集上的可视化检测效果如图7所示。每一行从左至右依次为基准模型检测结果、本文模型检测结果、基准模型的特征关注热力图以及本文模型的特征关注热力图，通过将检测结果与热力图并列呈现，可以直观比较两种模型在目标定位、注意力分布以及背景抑制方面的差异。

从图中的检测结果可见，在远距小尺度密集行人场景的第一行中，基准模型对位于车道尽头的小尺度行人以及左上护栏上的小尺度行人有明显漏检，而本文算法均能准确捕捉到上述被忽略的目标。对比其对应的热力图也能看出本文算法对正确行人特征的关注更加集中，同时对右侧背景噪声（道路施工警示柱）的响应也被显著抑制；在中距离遮挡的第二行图片中，Baseline虽能检测出近处行人，

但对被路边车辆严重遮挡的行人和道路尽头的小尺度行人仍响应不足，相比之下，本文算法不仅成功检测出上述被忽略的目标，其热力图还展现出对正确行人特征更加积极的响应；在第三行中，行人的尺度变化比前面两行更大。相比于RT-DETR，本文算法不仅能正确检测到近处行人目标，还捕捉到了之前被漏检的远处小尺度行人（中、右部分）。上述可视化结果表明，本文所提出的改进算法相较于RT-DETR基准模型能够更加精准地捕获远距离小尺度以及被遮挡的行人目标，其对应的注意力热力图在行人区域展现出更强、更明确的响应，改进策略强化了模型对正确行人特征的关注。

图8汇集了ACDC数据集中四类典型恶劣天气



(a) RT-DETR检测效果图 (b)FEBA-DETR检测效果图 (c)RT-DETR热力图 (d) FEBA-DETR热力图
((a) RT-DETR detection output; (b) FEBA-DETR detection output; (c) RT-DETR heatmap; (d) FEBA-DETR heatmap.)

图7　改进效果可视化对比图

Fig. 7　Visualization Comparison of the Improved Results

样本的检测结果,依次对应雨天、雪天、雾天以及夜晚的驾驶场景。每一类天气选取两组不同具体场景,分别展现RT-DETR基准模型与本文提出的改进模型在同一帧图像上的行人检测效果,以直观对比不同模型在复杂条件下的检测性能差异。

由图可以看出,基准模型在不同的天气条件下均有较为严重的漏检问题。在雨天场景(第一行),左侧对比组(场景一)显示雨幕和积水反光产生的高频噪声严重干扰了行人特征的有效识别,导致基准模型未能检测出两处中近景的行人目标;在右侧对比组中(场景二),基准模型漏检了远处小尺度行人。在雪和雾场景(第二、三行)中,大片积雪和浓雾导致图像整体对比度降低,对远距离小尺度行人带来强烈干扰,基准模型出现漏检问题。在夜晚场景(第四行)中,由于受弱光和局部强炫光干扰,左侧对比组(场景一)显示基准模型仅检测出近景行人,未能检测出中远暗处行人;右侧对比组(场景二)显示基准模型仅检测出右侧路口的行人,左侧较暗区域的行人被漏检。改进后的模型成功检测出上述漏检的行人目标,原因在于:(1)改进模型强化了基准模型网络结构本身的小目标检测能力和抗噪声能力;(2)改进模型引入频域子带数据增强来弥补由于上述外部环境干扰造成的细节特征退化和低对比度,使模型在训练中学习到更多去噪与细节增强后的恶劣天气样本。改进后的模型相比基准模型表现出明显的性能提升,验证了频域子带数据增强策略在应对雨、雾、雪和低光照等复杂环境下行人检测的有效性。

3　结　论

针对自动驾驶行人检测中小尺度目标检测困难、行人遮挡频繁、恶劣环境干扰以及全局信息不足等挑战,本文提出了一种结合多重频域增强策略与边界感知机制的行人检测算法FEBA-DETR。在数据增强阶段,本文使用基于频域子带处理的数据增强方法来应对恶劣环境干扰问题,使模型在训练过程中能学到更多去噪与细节增强后的恶劣天气样本。在特征提取阶段利用提出的FAM模块实现频域特征与空间域特征的初步融合,在FAIFI模块中进一步强化频空特征的深度关联,有效提高了模型对细微行人的检测能力和抗噪声能力。而边界感知机制则在特征融合阶段强化多尺度特征图中的边缘

(a) RT-DETR检测结果(场景一) (b)FEBA-DETR检测结果(场景一) (c)RT-DETR检测结果(场景二) (d)FEBA-DETR检测结果(场景二)

((a) RT-DETR detection output(Scene 1);(b) FEBA-DETR detection output(Scene 1);(c) RT-DETR detection output(Scene 2);(d) FEBA-DETR detection output(Scene 2))

图8　不利条件下改进效果对比图

Fig. 8　Comparison of Improved Performance under Adverse Conditions

细节信息,从而缓解对行人遮挡频繁的难题。从前述实验结果来看,FEBA-DETR相较于原模型在各项指标上均取得了提升,整体性能均优于当前主流的实时检测方法。验证了多重频域增强策略与边界感知机制在提高模型泛化能力与鲁棒性方面的有效性,为解决复杂环境下的行人检测提供了新的思路。

尽管本文的工作取得了上述进展,但模型在检测长尾分布下的稀有行人类别(如坐轮椅、推婴儿车、负重行走等特殊姿态)时仍存在局限。其本质原因在于稀有行人类别外观特征差异明显且训练样本稀缺,导致模型难以获得足够的特征表示。未来可结合少样本学习与自监督学习等方法,通过在行人数据集上预训练后快速适配稀有类别,从而更有效地捕获长尾类别的特异性,提升模型的鲁棒性与泛化能力。

参考文献(References)

Alfasly S, Chui C K, Jiang Q, Lu J and Xu C. 2024. An Effective Video Transformer With Synchronized Spatiotemporal and Spatial Self-Attention for Action Recognition.IEEE Transactions on Neural Networks and Learning Systems, 35(2): 2496 - 2509 [DOI:10.1109/TNNLS.2022.3190367]

Ang Fet al. 2024. DuAT: Dual-Aggregation Transformer Network for Medical Image Segmentation. Pattern Recognition and Computer Vision. PRCV 2023. Lecture Notes in Computer Science, 14429. Singapore:Springer[DOI:10.1007/978-981-99-8469-5_27]

Cao J, Pang Y, Xie J, Khan F S and Shao L. 2022. From Handcrafted to Deep Features for Pedestrian Detection: A Survey.IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9): 4913 - 4934 [DOI:10.1109/TPAMI.2021.3076733]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-End Object Detection with Transformers.Computer Vision - ECCV 2020. Cham: Springer: 213 - 229 [DOI:10.1007/978-3-030-58452-8_13]

Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, CA, USA: IEEE:886 - 893[DOI:10.1109/CVPR.2005.177]

Dollár P, Tu Z, Perona P and Belongie S J. 2009. Integral channel features.BMVC, 2(3): 5 [DOI: 10.5244/C.23.91.]

Dong C and Luo X S. 2021. Research on a pedestrian detection algorithm based on improved SSD network. Journal of Physics: Confer-

ence Series, 1802(3): 032073 [DOI: 10.1088/1742-659 6/1802/3/032073]

Fan X, Zhang Y, Lu Y and Wang H. 2024. PARFormer: Transformer-Based Multi-Task Network for Pedestrian Attribute Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 34(1): 411 - 423 [DOI: 10.1109/TCSVT.2023.3285411]

Felzenszwalb P F, Girshick R B, McAllester D and Ramanan D. 2010. Object Detection with Discriminatively Trained Part-Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627 - 1645[DOI:10.1109 /TPAMI.2 009.167]

Gao F, Leng J, Gan J and Gao X. 2023. Selecting Learnable Training Samples is All DETRs Need in Crowded Pedestrian Detect ion.Proceedings of the 31st ACM International Conference on Multimedia (MM '23). New York: ACM: 2714 - 2722 [DO I: 10.1145/3581783.3612189]

Ghari B, Tourani A, Shahbahrami A and Gaydadjiev G. 2024. Pedestrian detection in low-light conditions: A comprehensive survey. Image and Vision Computing, 148: 105106 [DOI:10.1016/j.imavis.2024.105106]

Gong An, Li Zhonghao, Liang Chenhong. 2023. NSPDet: real-time nearby-aware pedestrian detection algorithm for multi-scene surveillance at night. Journal of Image and Graphics, 28(09):2693-2705. (龚安, 李中浩, 梁辰宏. 2023. 夜间多场景的邻近感知实时行人检测算法. 中国图象图形学报,28(09):2693-2705[DOI:10.11834/jig.220834].

Huang S., Lu Z., Cun X., Yu Y., Zhou X. and Shen, X., 2024. DEIM: DETR with Improved Matching for Fast Convergenc e. arXiv preprint arXiv:2412.04234.[DOI:10.48550/arXiv.241 2.04234].

Khan A H, Nawaz M S and Dengel A. 2023. Localized Semantic Feature Mixers for Efficient Pedestrian Detection in Autonomous Driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 5476 - 5485 [DOI: 10.1109/CVPR52729.2023.00530]

Kong Y, Shang X and Jia S. 2024. Drone-DETR: Efficient Small Object Detection for Remote Sensing Image Using Enhanced RT-DETR Model.Sensors,24(17):5496[DOI:10.3390/s2417549 6]

Lee-Thorp J, Ainslie J, Eckstein I and Ontanon S. 2021. Fnet: Mixing tokens with fourier transforms. arXiv preprint arXiv: 2105.03824 [DOI:10.48550/arXiv.2105.03824].

Li Z, Kovachki N, Azizzadenesheli K, Liu B, Bhattacharya K, Stuart A and Anandkumar A. 2020. Fourier neural operator for parametric partial differential equations. arXiv preprint arXiv: 2010.08895. [DOI:10.48550/arXiv. 2010.08895]

Lin T Y, Dollár P, Girshick R, He K, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection. Procee dings of the IEEE Conference on Computer Vision and Pattern Recognition: 2117 - 2125. [DOI:10.1109/CVPR.2017.106]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: Single Shot MultiBox Detector. Computer Vision - ECCV 2016. Cham: Springer:21 - 37 [DOI :10.1007/ 9 78-3-319-46448-0_2]

Luo W, Li Y, Urtasun R and Zemel R. 2016. Understanding the effective receptive field in deep convolutional neural networks. Advances in Neural Information Processing Systems, 29: 4898-4906 [DOI:10.5555/3295222.3295343]

Ma S and Xu Y. 2023. Mpdiou: a loss for efficient and accurate bounding box regression. arXiv preprint arXiv: 2307.07662 [D OI: 10.48550/arXiv.2307.07662]

Peng Y, Li H, Wu P, Zhang Y, Sun X and Wu F. 2024. D-FINE: redefine regression Task in DETRs as Fine-grained distribution refinement. arXiv preprint arXiv: 2410.13842 [DOI: 10.48 550/ar Xiv. 2410.13842]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 779 - 788[DOI: 10.1109/CVPR.2016.91]

Ren S, He K, Girshick R and Sun J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Netwo rks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137 - 1149.[DOI:10.1109/TPAMI.2016.257 7031]

Sakaridis C, Dai D and Van Gool L. 2021. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. Proceedings of the IEEE/CVF Internat ional Conference on Computer Vision: 10765 - 10775 [DOI: 10.1109/ICCV48922.2021.01059]

Shorten C and Khoshgoftaar T M. 2019. A survey on Image Data Augmentation for Deep Learning. J Big Data, 6: 60 [DOI: 10.1 186/s40537-019-0197-0]

Tong Z, Chen Y, Xu Z and Yu R. 2023. Wise-IoU: bounding box regression loss with dynamic focusing mechanism. arXiv preprint arXiv:2301.10051[DOI: 10.48550/arXiv.2301.10051]

Wanjun Liu, Libing Dong, Haicheng Qu. Small-scale pedestrian detection based on improved R-FCN model[J]. Journal of image and graphics, 2021, 26(10): 2400-2410. (刘万军, 董利兵, 曲海成. 改进R-FCN模型的小尺度行人检测[J]. 中国图象图形学报,2021,26(10):2400-2410[DOI:10.11834/jig.2002 87]).

Wang C Y, Liao H Y M, Wu Y H, Chen P Y, Hsieh J W and Yeh I H. 2020. CSPNet: A new backbone that can enhance learning capability of CNN. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops: 390 - 391[DOI: 10.1109/CVPRW50498.2020.00203]

Wang K, Fu X, Huang Y, Cao C, Shi G, Zha Z-J. 2023. Generalized UAV Object Detection via Frequency Domain Disentanglement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023), pp. 1064-1073. DOI: 10.1109/CVPR52729.2023.00109

Wang X, Han T.X. and Yan S. 2009, September. An HOG-LBP human detector with partial occlusion handling. In 2009 IEEE 12th Inter-

national Conference on Computer Vision (pp. 32-39). IEEE [DOI: 10.1109/ICCV.2009.5459207].

Wang X, Zhang Y, Li Q and Chen X. 2022. RT-DETR: Real-Time Detection Transformer for Object Detection. IEEE Transactions on Intelligent Transportation Systems, 23(5): 1234-1243 [DOI: 10.1109/TITS.2022.3141592]

Wei R, He N and Yin X. 2020. YOLO-Person: Pedestrian Detection in Road Areas. Journal of Computer Engineering & Applications, 56(19):197 - 204). 魏润辰,何宁,尹晓杰.YOLO-Per son:道路区域行人检测[J].计算机工程与应用,2020,5 6(19):197-204). [DOI: 10.3778/j.issn.1002-8331.2004-0136]

Wu B, Wan A, Yue X, Jin P, Zhao S, Golmant N, Gholaminejad A, Gonzalez J and Keutzer K. 2018. Shift: A zero flop, zero parameter alternative to spatial convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 9127 - 9135. [DOI: 10.1109/CVPR.2018.00951]

Xinkai Xu, Yan Ma, Xu Qian, Yan Zhang. Scale-aware EfficientDet: real-time pedestrian detection algorithm for automated driving[J]. Journal of image and graphics, 2021, 26(1): 93-100. (徐歆恺, 马岩, 钱旭, 张龑. 自动驾驶场景的尺度感知实时行人检测[J]. 中国图象图形学报, 2021,26(1):93-100[DOI: 10.11834/jig.200445]).

Ye R, Gao Q, Qian Y, Sun J and Li T. 2024. Improved YOLOv8 and SAHI Model for the Collaborative Detection of Small Targets at the Micro Scale: A Case Study of Pest Detection in Tea. Agronomy, 14(5): 1034 [DOI:10.3390/agronomy14051034]

Zhang F, Panahi A and Gao G. 2023. FsaNet: Frequency Self-Attention for Semantic Segmentation. IEEE Transactions on Image Processing, 32: 4757 - 4772 [DOI:10.11 09/TI P.20 23. 3305090]

Zhang H, Xu C and Zhang S. 2023. Inner-iou: more effective intersection over union loss with auxiliary bounding box. arXiv preprint arXiv:2311.02877[DOI: 10.48550/arXiv.2311.02877]

Zhang H and Zhang S. 2024. Focaler-iou: More focused intersection over union loss. arXiv preprint arXiv:2401.10525[DOI: 10.48550/arXiv.2401.10525]

Zhang S, Wen L, Bian X, Lei Z and Li S Z. 2018. Occlusion-aware R-CNN: Detecting pedestrians in a crowd. Proceedings of the European Conference on Computer Vision (ECCV): 637 - 653[DOI: 10.1007/978-3-030-01219-9_39]

Zhang S, Benenson R and Schiele B. 2017. Citypersons: A diverse dataset for pedestrian detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 32 13 - 3221[DOI: 10.1109/CVPR.2017.474]

Zhang S, Xie Y, Wan J, Xia H, Li S Z and Guo G. 2020. WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild. IEEE Transactions on Multimedia, 22(2): 380 - 393 [DOI: 10.1109/TMM.2019.2929005]

Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, Liu Y and Chen J. 2024. Detrs beat yolos on real-time object detection. Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition:16965 - 16974[DOI:10.1109/CVPR52733. 2024 .01605]

Zhong Y, Li B, Tang L, Kuang S, Wu S, Ding S. 2022. Detecting Camouflaged Object in Frequency Domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), pp.4494-4503. [DOI: 10.1109/CVPR 52688.2022.00446]

作者简介

刘涛,男,副教授,主要研究方向为计算机视觉、智能数据处理。E-mail:303171362@qq.com

欧阳晖,男,硕士研究生,主要研究方向为图像处理、模式识别与人工智能。E-mail:472321742@stu.lntu.cn

高一萌,女,助教,主要研究方向为图像处理、模式识别与人工智能。E-mail:gymeng1031@163.com