Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(1), 267-276
Published Online January 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.151026

Hans汉斯

融合DFC注意力与特征剪枝的小目标三维检测与跟踪方法研究

杨鹏成

西华大学汽车与交通学院，四川 成都

收稿日期：2025年12月12日；录用日期：2026年1月5日；发布日期：2026年1月15日

摘　要

随着自动驾驶技术的逐步发展，车辆对环境中小目标的高效、准确感知(包括三维定位与持续跟踪)已成为其技术落地的关键。现有的视觉方法在目标尺度较小或发生部分遮挡时，由于像素信息稀疏、特征表达能力弱，三维检测精度和后续跟踪稳定性显著下降，难以满足复杂道路场景下的实时应用需求。针对上述问题，本文提出一种基于特征剪枝和可解耦全连接注意力(Decoupled Fully Connected Attention, DFC)机制的三维小目标检测与跟踪框架，在保持高实时性的同时提升小目标的三维检测与跟踪性能。首先，针对主干网络输出特征设计图像特征剪枝策略，对候选小目标区域进行深度挖掘以强化其表征能力。其次，在左右视图特征融合过程中引入硬件友好的DFC注意力机制，高效捕获长距离像素依赖并增强立体几何约束。最后，在检测网络输出的三维回归与分类结果基础上构建轻量级三维多目标跟踪模块，实现对小目标轨迹的准确关联与更新。为验证所提方法的有效性，我们在公开的基准数据集KITTI上进行了充分实验，与多种模型对比表明，该方法在小目标三维检测精度、实时性以及三维目标跟踪稳定性方面均取得了更优表现。

关键词

小目标检测，三维目标跟踪，自动驾驶，特征剪枝，注意力机制

Research on 3D Detection and Tracking Method for Small Objects Integrating DFC Attention and Feature Pruning

Pengcheng Yang

School of Automobile and Transportation, Xihua University, Chengdu Sichuan

Received: December 12, 2025; accepted: January 5, 2026; published: January 15, 2026

文章引用: 杨鹏成. 融合 DFC 注意力与特征剪枝的小目标三维检测与跟踪方法研究[J]. 人工智能与机器人研究, 2026, 15(1): 267-276. DOI: 10.12677/airr.2026.151026

Abstract

With the gradual development of autonomous driving technology, the efficient and accurate perception of small objects in the environment by vehicles (including three-dimensional positioning and continuous tracking) has become the key to the implementation of this technology. The existing visual methods, when the object scale is small or partial occlusion occurs, due to sparse pixel information and weak feature expression ability, the 3D detection accuracy and subsequent tracking stability significantly decline, making it difficult to meet the real-time application requirements in complex road scenarios. In response to the above problems, this paper proposes a 3D small object detection and tracking framework based on feature pruning and DFC (Decoupled Fully Connected) attention mechanism, which improves the 3D detection and tracking performance of small objects while maintaining high real-time performance. Firstly, an image feature pruning strategy is designed for the output features of the backbone network, and the candidate small object regions are deeply mined to enhance their representation ability. Secondly, in the process of fusion of left and right view features, a hardware-friendly DFC attention mechanism is introduced to efficiently capture long-distance pixel dependencies and enhance stereo geometric constraints. Finally, based on the three-dimensional regression and classification results output by the detection network, a lightweight three-dimensional multi-object tracking module is constructed to accurately associate and update the trajectories of small objects. To verify the effectiveness of the proposed method, we conducted thorough experiments on the public benchmark dataset KITTI. Comparisons with multiple models show that this method achieves superior performance in terms of the three-dimensional detection accuracy of small objects, real-time performance, and the stability of three-dimensional object tracking.

Keywords

Small Object Detection, 3D Object Tracking, Autonomous Driving, Feature Pruning, Attention Mechanism

1. 引言

面向自动驾驶的感知系统中，如何在车载算力受限条件下实现成本低、实时性高的三维环境理解，是车辆安全行驶的关键问题之一。对于远处的行人和小型车辆等小尺度目标，由于在图像中仅占据极少像素且深度与纹理信息不足，传统纯视觉三维检测在定位精度和召回率上都会明显下降，进而影响后续的跟踪稳定性。因此，面向小目标进行专门优化的三维检测与跟踪技术，正逐渐成为视觉感知研究的重要方向。

现有双目三维检测多采用伪 LiDAR 思路，如 Pseudo-LiDAR++ [1]先由双目图像估计深度，再转换为点云并交由点云检测器处理。ZoomNet [2]、DSGN [3]、OC Stereo [4]等则依赖额外的点云或深度图监督。此类方法虽能取得较高精度，但点云生成和三维卷积会带来大量的计算与存储开销，在远距离小目标场景下的深度恢复能力也有限。针对小目标本身，已有工作从数据增强、超分辨率、上下文建模、多尺度特征以及锚框和注意力设计等角度提升检测性能，但普遍存在对小目标提升有限或计算成本显著增加等问题。

近年来兴起的空间剪枝为轻量级视觉模型提供了新的思路。典型方法通过注意力掩码裁剪冗余 token，或利用背景体素稀疏性在点云与 BEV 特征上跳过无关区域[5] [6]。DSP 则专门面向点云小尺度目标三维检测，在无目标体素位置省略上采样操作以进一步降低开销[7]。这类方法在空间维度上有选择地计算特征，从而兼顾精度与效率，但大多聚焦于点云或 Transformer 结构，尚未充分挖掘双目图像浅层特征对小目标三维检测及跟踪的潜力。

基于上述分析，本文在双目视觉框架下提出面向小目标的动态特征剪枝与 DFC 注意力联合建模方法，构建端到端的三维检测与多目标跟踪一体化网络。我们在主干网络多尺度图像特征上引入动态特征剪枝策略，重点保留疑似小目标区域并抑制与任务无关的大目标及背景，实现对小目标浅层语义的高效提取。在左右视图融合阶段加入硬件友好的 DFC 注意力模块，以较低成本构建高质量视差代价体并增强长距离依赖建模。在此基础上，检测分支直接回归小目标三维框的尺度、位置、类别与朝向，同时设计轻量级三维多目标跟踪分支，在时间维度上对检测结果进行关联与轨迹更新。实验结果表明，该框架在小尺度目标的三维检测精度、推理速度以及连续跟踪稳定性方面均优于现有代表性方法。

2. 方法

2.1. 整体框架

整体网络框架如图 1 所示，可概括为检测分支与跟踪分支两个阶段。首先，将双目图像输入同一 ResNet [8]主干网络，在不同层级提取多尺度语义与空间特征。随后，左右视图在各个尺度上分别送入多尺度立体融合模块，完成立体匹配与特征交互，每一层融合单元内部都嵌入 DFC-Ghost 模块，以较低计算成本生成更加丰富的立体表征并增强长距离像素依赖。与此同时，来自左视图主干网络的多层特征被送入 DFP 模块，对与小目标相关的区域进行选择性放大、抑制冗余背景，得到小目标增强特征，再与左视图最高层特征进行拼接整合。之后，我们将左右视图融合得到的立体特征与增强后的左视图特征联合输入检测头，对三维边界框的类别、位置、尺寸以及朝向进行预测，得到稳定的三维检测结果。

在此基础上，利用检测结果构建时序输入，引入 AB3DMOT [9]作为统一的三维多目标跟踪基线，对连续帧中的候选目标进行数据关联与运动状态更新，形成了基于双目的小目标三维检测和统一多目标跟踪的一体化框架。一方面，DFC-Ghost 模块提升了立体几何信息的表达能力，为后端跟踪提供更加准确的三维位置与尺度估计。另一方面，DFP 模块显式强化 Hard 难度下小目标的特征响应，使得在同一 AB3DMOT 基线上，我们的检测–跟踪组合在轨迹连续性与实时性上均优于以 PointRCNN [10]等方法为前端的方案，这一点在第 3 节的实验对比中得到进一步验证。

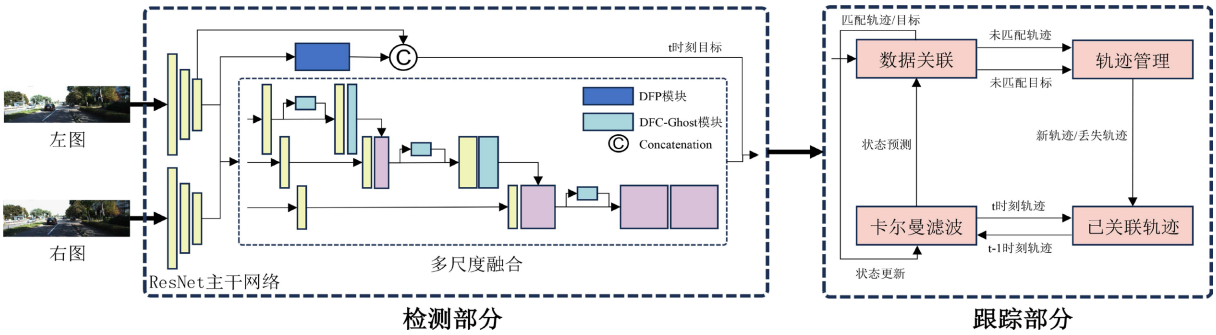

Figure 1. Overall network framework
图 1. 网络总体框架

2.2. DFC-Ghost 模块

在立体匹配中，传统做法常用特征拼接(Concatenation)构建代价体，右视图特征在通道维与左视图特征直接叠加，在每个视差位置将两者连接。若双目特征图尺寸为$[B,C,H,W]$，通过拼接得到的 cost volume 形状为$[B,2C,D,H,W]$。通道数随之翻倍，会显著增加后续三维卷积的计算量，使立体匹配难以满足实时要求。为提高效率，本文采用基于相关性的代价体构造方式，通过归一化点积计算左右特征在不同视差下的相似度，得到的 cost volume 维度为$[B,D,H,W]$，在保证匹配精度的前提下有效降低了计算开销。然而，此时三维特征通道数较少，再加上下采样操作带来的细节损失，立体信息不足，整体表示易偏向左视图，使三维框的预测精度受到限制。

为缓解上述问题，我们在代价体构造后引入 DFC-Ghost 模块，对特征进行增强并实现多尺度融合。Ghost 模块的核心思想是常规卷积产生的特征图中存在较多冗余分量，这些冗余对精度有益，但完全依赖标准卷积会带来过高的 FLOPs。因而先通过少量“核心特征”再用廉价线性变换生成冗余 ghost 特征，从而以更低成本获得近似等价的表示[11]。设输入特征为 $X\in\mathbb{R}^{c\times h\times w}$，其中 c 为通道数，h,w 为空间尺寸。标准卷积产生 n 个输出通道，可写为

$$Y = X * f + b \tag{1}$$

其中 $*$ 表示卷积，$f\in\mathbb{R}^{c\times k\times k\times n}$ 为卷积核，b 为偏置项，输出 $Y\in\mathbb{R}^{n\times h'\times w'}$。其计算量约为 $n\cdot h'\cdot w'\cdot c\cdot k\cdot k$。在 Ghost 结构中，先用一组较小的卷积核获取 m 个核心特征 $Y'\in\mathbb{R}^{m\times h'\times w'}$：

$$Y' = X * f' \tag{2}$$

其中 $f'\in\mathbb{R}^{c\times k\times k\times m}$。接着，对每个核心特征图 y'_i 应用若干个线性算子 $\Phi_{i,j}$，生成一组 ghost 特征 $\left\{y_{i,j}\right\}_{j=1}^{s}$，并将所有核心特征及其 ghost 特征拼接，得到总通道数 $n=m\cdot s$ 的输出 $Y=\left[y_{1,1},y_{1,2},\cdots,y_{m,s}\right]$。由于线性变换参数量远小于常规卷积，从而在接近的表达能力下显著降低了 FLOPs。

但核心特征 Y' 仍然由局部卷积产生，只能感知有限的感受野。为引入全局依赖，我们在 Ghost 结构中融入 DFC 注意力。设某一层特征 $Z\in\mathbb{R}^{h\times w\times c}$，可视作 $h\cdot w$ 个 token 组成的集合 $\left\{z_{ij}\right\}$。若直接用全连接层在二维平面上建模注意力，其复杂度约为 $O\left(h^2w^2\right)$，对实时推理不利。DFC 将二维全连接拆分为水平方向与垂直方向两个分支，分别学习对应方向上的长程依赖：

$$A_H = FC_H\left(Z\right),\ A_W = FC_W\left(Z\right) \tag{3}$$

其中 FC_H，FC_W 分别表示只在横向或纵向进行的线性变换，得到的注意力图再组合用于调制特征。这样计算复杂度可降为 $O\left(h^2w+hw^2\right)$，在保持较强建模能力的同时兼顾了速度。

综合上述思想，DFC-Ghost 模块构造为一个倒残差瓶颈结构，如图 2 所示，内部包含两个 Ghost 分支。第一个用于通道扩展，生成更丰富的中间表示。第二个在融合 DFC 注意力后进行通道压缩得到输出特征。具体实现中，输入特征分别进入 Ghost 分支和 DFC 分支，一条路径产生候选特征，另一条输出注意力图，两者做逐元素乘积后再次送入 Ghost 模块细化，最终与初始输入进行拼接，形成增强后的输出。

从复杂度与效果的角度看，DFC-Ghost 模块在相关性 cost volume 的基础上仅引入了少量线性变换与解耦全连接操作，相比直接叠加更深、更宽的三维卷积网络，其参数量和 FLOPs 增幅有限。在引入 DFC-Ghost 后，整体推理时间仍维持在 0.08 s/帧量级，却在 Easy、Moderate 与 Hard 三个难度上均带来了稳定的 3D AP 提升。这说明该模块在精度和速度权衡上是增益显著、代价可控的，既弥补了相关性 cost volume 通道不足及下采样带来的信息损失，又没有破坏系统的实时性，为后续小目标三维检测与多目标跟踪奠定了可靠的立体特征基础。

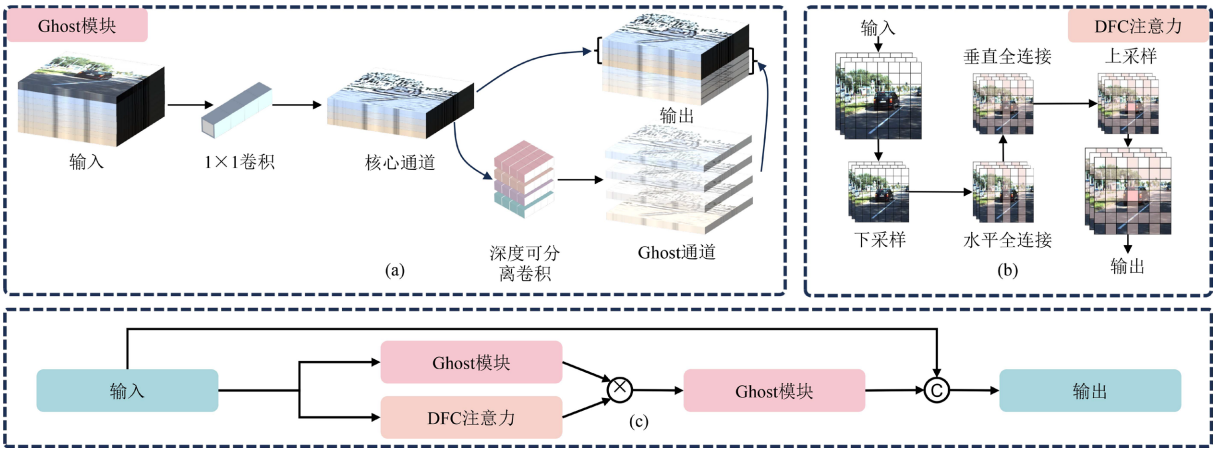

Figure 2. DFC-Ghost module structure
图 2. DFC-Ghost 模块结构

2.3. DPF 模块

在整体网络设计中，DFP 模块只作用于左视图的高层特征，而不直接处理右视图的深层特征。原因在于，右图与左图之间天然存在视差偏移，只有经过多尺度立体融合模块后得到的双目联合特征，才能在同一坐标系中与左图的语义空间精确对齐。如果将尚未对齐的右视图高层特征直接拼接或叠加到检测分支，极易引入额外噪声，削弱三维框回归与分类的稳定性。同时，若将左右视图的深层特征全部与立体融合特征一起连接，通道维度会急剧膨胀，既显著提高计算量和显存开销，也会使输出特征在通道分布上过度偏向原始 RGB 纹理，从而压制对精细视差信息和立体结构的建模能力。

为提升左图中小目标的表达效果，本文设计了动态特征剪枝模块 DFP，其结构如图 3 所示。该模块通过可学习的剪枝比例 r_i 以及概率排序相结合的掩码机制，在特征图上自动选出响应度较高、潜在包含小目标的位置；同时，将低分辨率特征中携带的全局上下文信息反向投影到高分辨率尺度，在高分辨率空间内补充小目标的语义线索，缓解多次下采样带来的信息缺失。大面积与任务无关的区域则被显式剔除，仅在关键位置(疑似小目标处)保留特征，实现对小目标表示的强化与整体计算开销的压缩。

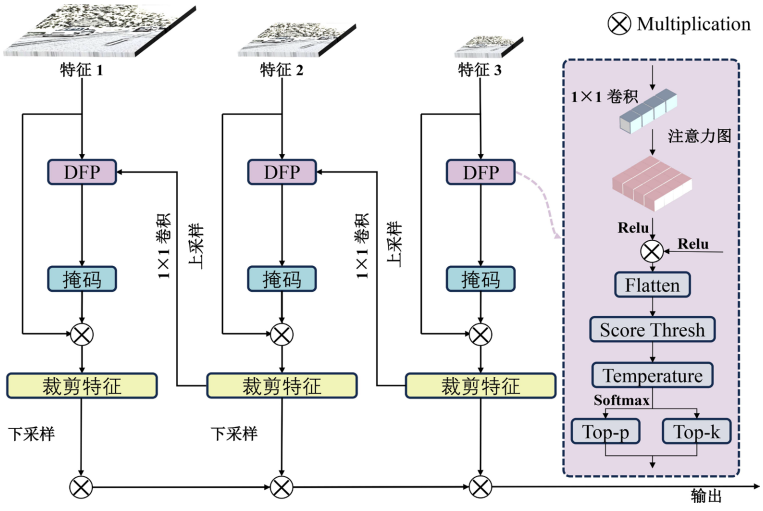

Figure 3. DFP module structure
图 3. DFP 模块结构

DFP 的输入由三个尺度的特征图$\{F_i\}$组成。对于每个尺度i，引入一个可学习标量，并通过 sigmoid 函数映射得到保留比例，再据此计算需要保留的元素个数：

$$K_i = H_i W_i \times r_i \tag{4}$$

其中H_i，W_i分别表示第i层特征F_i的高和宽。由于K_i是动态学习得到的，网络在训练过程中可以自适应调节不同尺度的保留密度。对于小目标分布更密集的高分辨率层，模块会倾向于学习更大的保留比例，以避免对细粒度结构过度剪枝；而在低分辨率层中，则可适当减小K_i，从而减少冗余计算。

当特征F_i进入 DFP 模块后，首先通过1×1卷积进行初始打分，生成第i层的特征评分图：

$$P_i = \sigma\left(Conv_{1\times1}\left(F_i\right)\right) \tag{5}$$

其中$\sigma(\cdot)$为非线性激活函数。经过这一步，高响应区域更集中于小目标或其他关键结构。当$i \neq 3$时，还会进行跨尺度信息交互。将更低分辨率层中已被剪枝的特征上采样到与当前尺度一致，经1×1卷积投影后得到跨尺度特征，再与P_i融合，形成融合得分图S_i。这种跨尺度投影能够把低分辨率的全局语义补充到高分辨率局部响应中，使小目标的显著性更加清晰。对每个融合得分$S_i\left(i \neq 3\right)$，我们先将其展平成一维向量，对低响应值进行 score threshold，抑制明显噪声，随后施加带温度参数的 softmax：

$$\tilde{s}_i = \mathrm{softmax}\left(\frac{S_i}{T}\right) \tag{6}$$

其中T为温度系数，用于控制分布的平滑程度。接下来，利用 Top-P 掩码m_p保留累积概率不超过p_{thresh}的位置，同时通过 Top-K 掩码m_k强制选出前K_i个最高响应点。二者合并后再重塑回原始空间尺寸，得到最终的二值掩码M_i，并用于特征筛选：

$$\hat{F}_i = F_i \odot M_i \tag{7}$$

其中$\odot$表示逐元素乘积操作。该混合策略一方面保证最显著的K_i个关键位置(通常对应小目标或其边缘区域)不会被误删，另一方面又能在剩余区域按概率质量适度保留部分上下文，为目标与背景的区分提供辅助信息。最后，对于$i \neq 3$的各尺度输出$\hat{F}_i$，我们按从高分辨率到低分辨率的顺序依次进行下采样与拼接，将多尺度的小目标增强特征在通道维度上融合，得到最终的增强特征，并送入后续的三维检测与跟踪分支。这使得网络在保持整体效率的同时，显著提升了对远距离和小尺度目标的三维感知能力。

需要强调的是，DFP 模块与 KITTI 数据集中 Hard 难度样本的特性高度契合。此类样本往往目标尺寸较小、遮挡严重、截断率高，若采用均匀计算策略，模型容易被大面积背景与近处大目标影响。通过动态剪枝与跨尺度投影，DFP 在空间维度上形成了高分辨率细粒度加上低分辨率全局语义的互补结构，使网络在有限计算预算下，将更多算力集中到疑似小目标区域。第 3 节的对比结果表明，Hard 难度下的 AP 提升幅度显著高于 Easy、Moderate 场景，印证了 DFP 对小目标检测的针对性贡献，也为后续在时序维度上保持小目标轨迹稳定提供了更强的单帧特征基础。

3. 实验验证

3.1. 基准数据集

在实验中，模型的检测部分基于 KITTI Object Detection Benchmark [12]进行训练与评估。该数据集为车载场景构建，提供同步采集的 RGB 图像、激光雷达点云、三维标注真值以及传感器标定文件等多模态信息，能够较为完整地刻画道路环境中的三维结构。官方标注的语义类别包含‘Car’、‘Van’、‘Truck’、‘Pedestrian’、‘Person_sitting’、‘Cyclist’、‘Tram’、‘Misc’和‘DontCare’等多种交通参与者及忽略区域。参考 baseline model 中的设置，本文仅选取“汽车(Car)”和“行人(Pedestrian)”两个类别参

与训练和测试。KITTI 三维目标检测基准中共提供 7481 帧用于训练，7518 帧作为测试。按照 Chen 等人[13]的划分方案，本文将 7481 帧训练数据进一步分为 3712 帧训练子集与 3769 帧验证子集，便于与已有方法进行公平对比。

为了验证所提出检测–跟踪一体化框架在时序场景中的有效性，本文在完成三维检测网络训练后，又引入 KITTI Multi-object Tracking Benchmark 对多目标跟踪分支进行评估。与检测基准类似，跟踪数据集同样由车载传感器采集的连续帧图像构成，并提供逐帧的二维/三维边界框标注以及跨帧的一致目标 ID，可同时服务于检测质量与轨迹连贯性的联合分析。在跟踪实验中，我们保持与检测部分一致的类别选择，主要关注行人和车辆在连续帧中的三维位置与身份保持性能。通过在 KITTI 检测和跟踪两个数据集上的联合实验，我们能够系统性地评估所提方法在单帧三维几何预测与跨帧时序关联两方面的性能表现。

3.2. 小目标检测实验

在检测实验部分，本文主要侧重分析在不同难度划分下的检测表现与小目标感知能力。评价指标采用官方 3D AP (Average Precision)度量，在 Cars 类别上使用 IoU = 0.7 阈值，在 Pedestrian 类别上采用 IoU = 0.5 阈值。KITTI 官方根据目标高度、遮挡程度和截断比例将样本划分为 Easy、Moderate 和 Hard 三个等级，其中 Hard 更倾向于尺寸较小且被严重遮挡或部分截断的样本，因此在本文中可视作小目标场景的代表，用于验证网络对于小目标的检测性能。

首先从车辆检测的结果来看，如表 1 所示，在仅使用双目图像输入的前提下，我们将所提网络与 StereoRCNN [14]、YOLOStereo3D [15]等代表性方法进行对比。相较 YOLOStereo3D，所提网络在 Car 类别的 Easy、Moderate、Hard 三个难度下 3D AP 分别提升约+2.36%、+0.55%和+0.54%，说明 DFP 与 DFC-Ghost 的结合能够在不牺牲速度的前提下，进一步提升双目三维几何回归的精度。同时与一系列依赖点云数据或预训练视差估计模块的方案 Pseudo-LiDAR、OC Stereo、ZoomNet、Disp R-CNN、Pseudo-LiDAR++、DSGN 相比，我们在 Hard 难度上取得相近甚至更优的 AP，同时推理时间控制在约 0.08 s/帧，远低于部分基于点云方法 0.3~0.6 s/帧的水平，体现出明显的实时性优势。

在行人检测任务上，我们在同一验证集划分下对 Pedestrian 类别进行了进一步评估，如表 2 所示。实验结果表明，所提网络在 Easy、Moderate 与 Hard 三个难度上均超过对比方法，尤其是 Hard 场景中提升更为突出。相较 YOLOStereo3D，三种难度下的 3D AP 分别提升约+2.08%、+0.96%和+1.79%。这说明 DFP 模块有效缓解了小尺度行人特征易被下采样和背景淹没的问题，而 DFC-Ghost 生成的多尺度立体特征也为行人三维框的稳定回归提供了更充足的深度信息。综合车辆和行人两个类别的结果可以看出，本方法在整体精度、对小目标的敏感度以及实时性之间取得了较为理想的平衡，为自动驾驶场景中的高效三维检测与后续三维跟踪提供了可靠基础。

Table 1. Comparison of 3D target detection results for car categories by different methods
表 1. 不同方法的车辆类别 3D 目标检测结果对比

方法	Easy	Moderate	Hard	Time
基于点云				
Pseudo-LiDAR	61.90	45.30	39.00	0.40 s
OC Stereo	64.07	48.34	40.39	0.35 s
ZoomNet	62.96	50.47	43.63	0.35 s
Disp R-CNN	64.29	47.73	40.11	0.42 s

续表

Pseudo-LiDAR++	63.20	46.80	39.80	0.40 s
DSGN	72.31	54.27	47.71	0.67 s
基于双目视觉				
StereoRCNN	54.11	36.69	31.07	0.30 s
YOLOStereo3D	70.06	46.58	35.53	0.08 s
Ours	72.42	47.13	36.07	0.08 s

Table 2. Comparison of 3D target detection results for pedestrian categories by different methods
表 2. 不同方法的行人类别 3D 目标检测结果对比

方法	Easy	Moderate	Hard	Time
基于点云				
Pseudo-LiDAR	33.80	27.40	24.00	0.40 s
OC Stereo	34.80	29.05	28.06	0.35 s
基于双目视觉				
YOLOStereo3D	37.46	29.04	23.25	0.08 s
Ours	39.54	30.00	25.04	0.08 s

3.3. 多目标跟踪实验

在多目标跟踪实验中，本文沿用检测部分对小目标的定义，将满足 KITTI 数据集中 Hard 难度条件的车辆和行人统一视作小目标场景。在相同的 AB3DMOT 跟踪基线上，我们分别采用 PointRCNN 与本文提出的双目检测网络作为前端，保持跟踪参数与评价流程一致，以突出前端三维检测质量对整体跟踪表现的影响，可视化结果如图 4 所示。

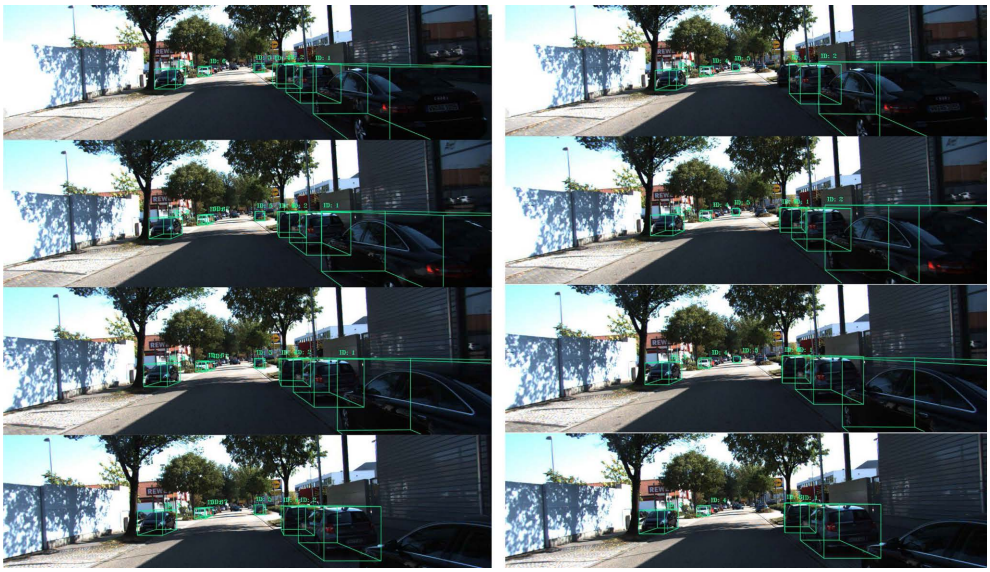

Figure 4. Visual comparison of multi-object tracking results
图 4. 多目标跟踪结果可视化对比

从可视化对比可以观察到，在远距离或被部分遮挡的小目标上，基于 PointRCNN 的方案更容易出现轨迹片段化、ID 频繁切换以及三维边界框抖动等问题。而采用本文检测结果作为输入时，Hard 难度下的车辆与行人轨迹显著更为连贯，三维框在深度方向和横向位置上与真实轨迹贴合度更高，尤其在车流密集或行人聚集的复杂场景中，小目标的跟踪稳定性提升更为明显。此外，在维持更高跟踪精度和轨迹连贯性的同时，整体推理速度也由基于 PointRCNN 的约 207 FPS 提升至约 270 FPS，证明所提出的小目标三维检测与跟踪框架在同一跟踪基线下实现了精度提升和速度加速的双重收益，进一步印证了方法设计的合理性。

4. 总结

本文围绕自动驾驶场景下远距行人和小型车辆等小目标难以稳定感知的问题，构建了一个面向双目视觉的三维检测与多目标跟踪一体化框架。方法在左视图主干特征上引入动态特征剪枝策略，显式突出小尺度区域并压缩冗余背景。在立体匹配阶段通过相关性构造代价体，并结合 DFC-Ghost 结构高效生成多尺度立体特征，在较低计算开销下增强长距离像素关联与深度表征。基于该检测网络，进一步接入 AB3DMOT 三维多目标跟踪模块，对时序帧中的车辆与行人轨迹进行关联与更新。实验结果表明，所提方法在 KITTI 检测基准上相较现有双目与点云方法，在小目标上取得了更优的 3D AP 与更好的实时性平衡。在同一跟踪基线上，相比以 PointRCNN 为前端的方案，小目标轨迹更加连续、定位更加精确，整体速度由约 207 FPS 提升至约 270 FPS。

未来工作中，我们计划从以下几个方向进一步扩展和完善本文方法：第一，将框架迁移到 nuScenes、Waymo Open Dataset 等包含更多传感器形态与复杂交通参与者类型的大规模数据集上，系统评估在多天气、多光照和多场景条件下的泛化能力；第二，探索与 BEV 表示、Transformer 结构等新型三维感知架构的结合方式，使 DFP 与 DFC-Ghost 可以模块化集成到更多前端检测器中；第三，在检测与跟踪联合优化方面，引入端到端训练策略或联合损失设计，增强时空一致性建模，从而进一步提升在密集车流、夜间驾驶、恶劣天气等极端场景下的小目标稳健感知能力。这些方向有望推动本文方法从研究原型走向更大规模、更加复杂的实际自动驾驶应用场景。

参考文献

[1] You, Y., Wang, Y., Chao, W.L., *et al*. (2019) Pseudo-Lidar++: Accurate Depth for 3D Object Detection in Autonomous Driving.

[2] Pang, Y., Zhao, X., Xiang, T., Zhang, L. and Lu, H. (2022) Zoom in and out: A Mixed-Scale Triplet Network for Camouflaged Object Detection. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 2160-2170. https://doi.org/10.1109/cvpr52688.2022.00220

[3] Chen, Y., Liu, S., Shen, X. and Jia, J. (2020) DSGN: Deep Stereo Geometry Network for 3D Object Detection. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 12536-12545. https://doi.org/10.1109/cvpr42600.2020.01255

[4] Pon, A.D., Ku, J., Li, C. and Waslander, S.L. (2020) Object-Centric Stereo Matching for 3D Object Detection. 2020 *IEEE International Conference on Robotics and Automation* (*ICRA*), Paris, 31 May-31 August 2020, 8383-8389. https://doi.org/10.1109/icra40945.2020.9196660

[5] Rao, Y., Zhao, W., Liu, B., *et al*. (2021) Dynamicvit: Efficient Vision Transformers with Dynamic Token Sparsification. 34*th Conference on Neural Information Processing Systems* (*NeurIPS* 2020), 6-12 December 2020, 13937-13949.

[6] Zhao, T., Ning, X., Hong, K., Qiu, Z., Lu, P., Zhao, Y., *et al*. (2023) Ada3d: Exploiting the Spatial Redundancy with Adaptive Inference for Efficient 3D Object Detection. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 2-3 October 2023, 17728-17738. https://doi.org/10.1109/iccv51070.2023.01625

[7] Xu, X., Sun, Z., Wang, Z., Liu, H., Zhou, J. and Lu, J. (2024) DSPDet3D: 3D Small Object Detection with Dynamic Spatial Pruning. In: *European Conference on Computer Vision*, Springer, 355-373. https://doi.org/10.1007/978-3-031-73390-1_21

[8] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[9] Weng, X., Wang, J., Held, D., *et al*. (2020) Ab3dmot: A Baseline for 3d Multi-Object Tracking and New Evaluation Metrics.

[10] Shi, S., Wang, X. and Li, H. (2019) Pointrcnn: 3D Object Proposal Generation and Detection from Point Cloud. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 16-21 June 2019, 770-779. https://doi.org/10.1109/cvpr.2019.00086

[11] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C. and Xu, C. (2020) Ghostnet: More Features from Cheap Operations. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 1580-1589. https://doi.org/10.1109/cvpr42600.2020.00165

[12] Geiger, A., Lenz, P., Stiller, C. and Urtasun, R. (2013) Vision Meets Robotics: The KITTI Dataset. *The International Journal of Robotics Research*, **32**, 1231-1237. https://doi.org/10.1177/0278364913491297

[13] Chen, X., Kundu, K., Zhu, Y., *et al*. (2015) 3D Object Proposals for Accurate Object Class Detection. *Proceedings of the* 29*th International Conference on Neural Information Processing Systems*, Volume 1, 424-432.

[14] Li, P., Chen, X. and Shen, S. (2019) Stereo R-CNN Based 3D Object Detection for Autonomous Driving. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 16-21 June 2019, 7644-7652. https://doi.org/10.1109/cvpr.2019.00783

[15] Liu, Y., Wang, L. and Liu, M. (2021) Yolostereo3d: A Step Back to 2D for Efficient Stereo 3D Detection. 2021 *IEEE International Conference on Robotics and Automation* (*ICRA*), Xi'an, 30 May-5 June 2021, 13018-13024. https://doi.org/10.1109/icra48506.2021.9561423