·智能汽车全栈环境感知技术专题·

基于改进BEVFormer的多视角图像目标检测算法*

李紫薇　艾子豪

（吉林动画学院，长春 130012）

【摘要】为了提升自动驾驶多摄像头视觉感知中鸟瞰视角（BEV）三维目标检测的精度与鲁棒性，提出一种基于改进BEVFormer的多视角图像目标检测方法。在图像特征层引入透视区域级辅助监督，通过BEV标注图像视角下的投影约束，引导骨干网络学习三维空间定位的局部细粒度特征；在BEV表征层构建实例级一致性约束，基于对比学习机制显式约束同一物理目标在不同视角条件下的BEV特征一致性，缓解视图变换引发的实例级特征漂移。基于nuScenes数据集对所提出方法进行训练与评估，并与多种典型多视角图像三维目标检测算法进行对比分析。试验结果表明：相较于BEVFormer基线模型，所提出方法在平均精度（mAP）和综合检测得分（NDS）指标上分别提升1.6百分点和0.4百分点，平均定位误差下降8.7百分点，验证了该方法在多视角图像BEV感知任务中的有效性。

主题词：多视角图像　鸟瞰视角感知　BEVFormer　三维目标检测

中图分类号：U495　　**文献标志码：**A　　**DOI:** 10.19620/j.cnki.1000-3703.20260019

Multi-View Image Object Detection Algorithm Based on Improved BEVFormer

Li Ziwei, Ai Zihao

（Jilin Animation Institute, Changchun 130012）

【Abstract】To improve the accuracy and robustness of 3D object detection from a Bird's-Eye View (BEV) in multi-camera visual perception for autonomous driving, this paper proposes a multi-view image object detection method based on an improved BEVFormer. To introduce perspective-based regional auxiliary supervision at the image feature level, using BEV annotations to impose projection constraints on the image view, thereby guiding the backbone network to learn fine-grained local features for 3D spatial localization. This paper establishes instance-level consistency constraints, at the BEV representation level and explicitly enforces the consistency of BEV features for the same physical object across different viewpoints based on a contrastive learning mechanism, thereby mitigating instance-level feature drift caused by view transformations. Furthermore, this paper trains and evaluates the proposed method using the nuScenes dataset and conduct a comparative analysis with various typical multi-view 3D object detection algorithms. The experimental results show that, compared to the BEVFormer baseline model, the proposed method achieves improvements of 1.6 percentage points and 0.4 percentage points in mean Average Precision (mAP) and the Normalized Detection Score (NDS), respectively, while reducing the average localization error by 8.7 percentage points, thereby validating the effectiveness of this method for the BEV detection task using multi-view images.

Key words: Multi-view images, Bird's-Eye View (BEV), BEVFormer, 3D object detection

【引用格式】李紫薇, 艾子豪. 基于改进BEVFormer的多视角图像目标检测算法[J]. 汽车技术, 2026(5): 15-22.

LI Z W, AI Z H. Multi-View Image Object Detection Algorithm Based on Improved BEVFormer[J]. Automobile Technology, 2026(5): 15-22.

1　前言

鸟瞰视角（Bird's-Eye View，BEV）感知通过将多视角传感器的数据统一映射到自顶向下的空间表示中，能够有效缓解因视角差异引发的几何不一致问题，逐渐成为自动驾驶与机器人系统中的关键技术之一[1]。该技术在三维目标检测、语义分割、轨迹预测及运动规划等任务中具有良好的通用性与实用价值[2-4]，为环境建模

*基金项目：吉林省高教科研课题（JGJX25C205）。

提供结构化且尺度一致的表达形式。

典型的BEV感知模型通常由图像特征提取主干、BEV编码器以及任务相关预测头构成，在统一空间坐标系下充分融合多视角与时序信息，服务于下游三维感知任务[5]。BEV感知建模主要从表示构建范式、几何建模与工程效率等方面展开。Liu等[6]通过将检测范式直接迁移至BEV空间，结合数据增强与后处理策略，使模型在精度与效率间达到较好平衡。Wang等[7]基于3D查询自上而下的建模方式，降低了逐像素深度估计产生的误差累积，简化了多视角到BEV的映射过程。针对深度估计不确定性问题，Wang等[8]通过概率化实例深度表示与几何关系建模，显著提升了图像驱动三维检测的可靠性。FCOS3D将二维检测范式迁移至三维任务，无需额外深度预测模块即可实现端到端3D检测，大幅降低了纯视觉三维目标检测的建模难度与计算复杂度[9]。Li等[10]设计BEVFormer框架，通过时空Transformer统一跨视角与时序信息，在多摄像头条件下实现了BEV表征学习，但复杂交通场景中的视图变换误差和深度不确定性仍存在实例级表征不一致。在工程部署层面，Dong等[11]从硬件延迟出发，通过设计硬件感知的网络与算子，实现了精度与推理效率的协同优化，但未关注多视图表征一致性，且在复杂场景中模型的鲁棒性较差。

在表示学习领域，虽然在对比学习中引入BEV感知框架能够增强特征判别性与一致性，通过额外的表征约束弥补传统检测损失在特征学习方面的不足[12]，然而，图像级对比学习策略直接应用于BEV感知任务无法获得稳定增益，甚至在部分数据集上使检测性能下降[13]。主要原因在于自动驾驶感知数据集的目标类别与场景多样性相对有限，难以构建具有充分判别性的正、负样本分布。同时，图像级对比学习侧重全局一致性，难以刻画BEV感知中多目标、复杂背景条件下的对象级与区域级细粒度差异，从而限制了其对三维检测任务的实际贡献[14-15]。

因此，本文以BEVFormer为基础框架，从表示学习角度引入与BEV构建过程相适配的对比监督机制，以提升多视角图像BEV表征的判别性与稳定性。在训练阶段，分别在BEV编码器输出层和二维图像骨干网络的中、高层特征上，引入实例级与透视区域级对比监督，前者强化实例级一致性，后者增强区域级判别性，从而共同提升BEV感知模型的特征表达能力。

2　改进BEVFormer的目标检测算法

2.1　总体框架

本文构建的端到端感知总体框架如图1所示，系统以多视角同步图像为输入，先由共享权重的2D骨干网络提取多尺度特征；再通过基于查询的空间交叉注意力将各视角特征投影并聚合到统一的BEV查询空间；随后利用时序自注意力融合历史与当前BEV信息，构建对动态场景具有鲁棒性的BEV表征；最终由BEV检测头预测目标类别与三维框参数。

针对模型在复杂交通场景（如遮挡、远距小目标等）的表征局限，本文在训练阶段引入2类针对性辅助约束：一是在BEV编码器输出处实施实例级一致性约束，以强化同一物理目标在不同视角或不同数据增强下的BEV表征一致性；二是在2D骨干的中高层对由BEV标注投影得到的图像区域施加透视区域级一致性约束，以引导骨干网络增强对三维定位关键区域的响应。

2.2　符号与基础定义

假定t时刻的多视角RGB输入集合为$\boldsymbol{J}_t=\{\boldsymbol{I}_t^k\}_{k=1}^{N_v}$，其中，$N_v$为摄像头数量，$\boldsymbol{I}_t^k$为$t$时刻第$k$个摄像头采集的

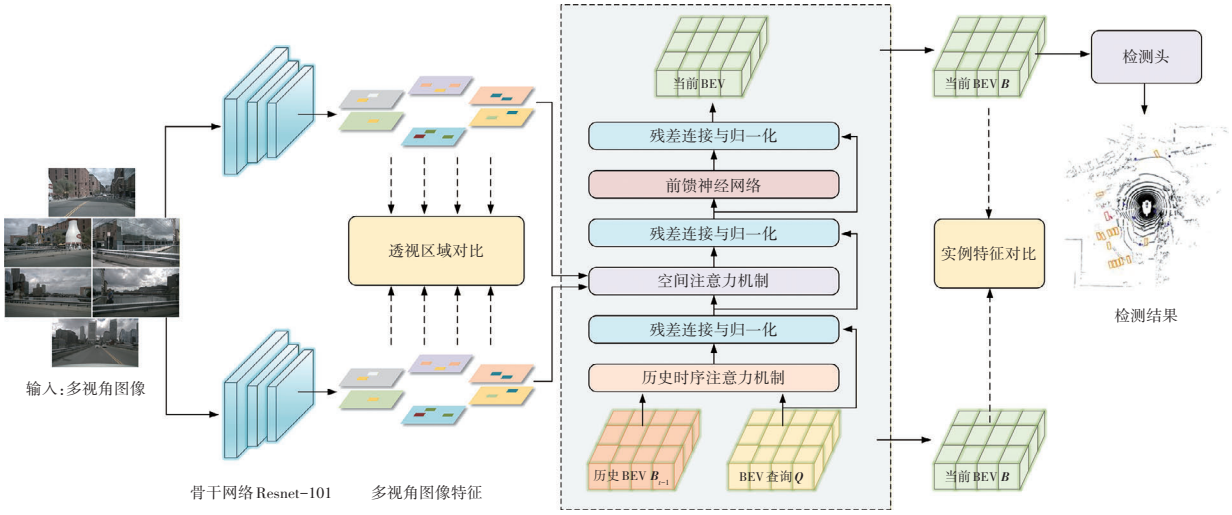


图1　总体框架

原始RGB图像。2D骨干网络在尺度层的输出为 $\boldsymbol{P}_j=\{\boldsymbol{P}_j^k\}_{k=1}^{N_v}$，多尺度集合为 $\boldsymbol{P}=\{\boldsymbol{P}_j\}_{j=1}^{N_1}$，$N_1$ 为骨干网络输出特征尺度层数。BEV网格设为 $H\times W$，特征张量为 $\boldsymbol{B}\in\mathbf{R}^{H\times W\times C}$，其中，$H$、$W$、$C$ 分别为BEV网格的高度、宽度和特征通道数。第 $p=(x,y)$ 个查询向量为 $\boldsymbol{Q}_p$，$\boldsymbol{Q}\in\mathbf{R}^{H\times W\times C}$。

将BEV查询映射至平面，利用预定义高度锚集合 $\{z_r'\}_{r=1}^{N_{\text{ref}}}$ 与各相机的投影矩阵 $\boldsymbol{T}_k\in\mathbf{R}^{3\times4}$，其中，$N_{\text{ref}}$ 为预定义高度锚点总数。BEV网格中位置 p 的世界坐标为：

$$\begin{cases}x'=\left(x-\dfrac{W}{2}\right)\cdot s\\ y'=\left(y-\dfrac{H}{2}\right)\cdot s\end{cases}\tag{1}$$

式中：x、y 为BEV的网格坐标，x'、y' 为对应世界坐标，s 为BEV网格分辨率。

对应高度锚 z_r' 的3D点 (x',y',z_r') 在相机 k 上的投影满足以下变换：

$$z_{p,k,r}\begin{bmatrix}x_{p,k,r}\\ y_{p,k,r}\\ 1\end{bmatrix}=\boldsymbol{T}_k\begin{bmatrix}x'\\ y'\\ z_r'\\ 1\end{bmatrix}\tag{2}$$

BEVFormer在空间域的特征采样采用可变形注意力（Deformable Attention，DA）：

$$DA(\boldsymbol{q},p,X)=\sum_{h=1}^{H_{\text{head}}}\boldsymbol{W}_h\left(\sum_{m=1}^{M}A_{h,m}\,\boldsymbol{W}_h'X(p+\Delta p_{h,m})\right)\tag{3}$$

式中：$\boldsymbol{q}$ 为查询向量，$X()$ 为输入特征的双线性插值采样，$A_{h,m}$、$\Delta p_{h,m}$ 分别为注意力权重与可学习偏移，$\boldsymbol{W}_h$、$\boldsymbol{W}_h'$ 为头部投影矩阵。

多视图特征 $\{\boldsymbol{F}_t^k\}$ 上的空间交叉注意力（Spatial Cross-Attention，SCA）为：

$$SCA(\boldsymbol{Q}_p,\{\boldsymbol{F}_t^k\})=\frac{1}{|V_{\text{hit}}(p)|}\sum_{k\in V_{\text{hit}}(p)}\sum_{r=1}^{N_{\text{ref}}}DA(\boldsymbol{Q}_p,P(p,k,r),\boldsymbol{F}_t^k)\tag{4}$$

式中：$V_{\text{hit}}(p)$ 为投影后命中的视图集合。

对经自运动补偿的历史BEV特征 $\boldsymbol{B}'_{t-1}$ 与当前目标查询进行时序交互，计算时序自注意力（Temporal Self-Attention，TSA）：

$$TSA(\boldsymbol{Q}_p,\{\boldsymbol{Q},\boldsymbol{B}'_{t-1}\})=\sum_{V\in\{\boldsymbol{Q},\boldsymbol{B}'_{t-1}\}}DA(\boldsymbol{Q}_p,p,V)\tag{5}$$

式中：V 为注意力计算的特征值集合，$\boldsymbol{Q}$ 为当前目标查询特征，$\boldsymbol{B}'_{t-1}$ 为与自运动补偿后的历史BEV特征。

2.3 实例级BEV对比学习

实例级对比学习以BEV空间为介质，约束同一物理实例在不同视角或不同数据增强下的表征一致性，从而弥补检测损失在深层Transformer反向传播时的梯度稀释问题。从BEVFormer编码层 l 的输出 $\boldsymbol{B}^{(l)}$ 中，提取实例描述并在对比空间计算判别损失，采用多层并行监督覆盖不同语义尺度。

已知当前帧的实例集合 $A=\{A_i\}_{i=1}^{n}$，其中，每个实例在BEV平面上以矩形或多边形表示，n 为当前帧的实例目标总数。对第 l 层输出 $\boldsymbol{B}^{(l)}\in\mathbf{R}^{H\times W\times C}$，针对BEV平面上的实例 A_i，通过RoIAlign完成特征对齐与池化，得到固定维度的特征图，再经全局池化映射为 C 维向量：

$$\boldsymbol{f}_i^{(l)}=f_{\text{Pooling}}(f_{\text{RoIAlign}}(\boldsymbol{B}^{(l)},A_i)),\boldsymbol{f}_i^{(l)}\in\mathbf{R}^C\tag{6}$$

对 $\boldsymbol{f}_i^{(l)}$ 进行 L_2 归一化处理，得到 $\bar{\boldsymbol{f}}_i^{(l)}=\boldsymbol{f}_i^{(l)}/\|\boldsymbol{f}_i^{(l)}\|_2$。采用数据增强或由动量编码器生成的BEV对 $\boldsymbol{B}'^{(l)}$ 提取对应归一化特征 $\bar{\boldsymbol{f}}_i'^{(l)}$。单层实例对比损失采用温度化的交叉熵（Normalized Temperature-Scaled Cross Entropy，NT-Xent）：

$$\begin{cases}L_{\text{in}}^{(l)}=\dfrac{1}{n}\sum_{i=1}^{n}l_i^{(l)}\\ l_i^{(l)}=-\log\dfrac{\exp\left(\dfrac{\bar{\boldsymbol{f}}_i^{(l)\mathrm{T}}\bar{\boldsymbol{f}}_i'^{(l)}}{\tau}\right)}{\sum\limits_{j=1,j\neq i}^{n}\exp\left(\dfrac{\bar{\boldsymbol{f}}_i^{(l)\mathrm{T}}\bar{\boldsymbol{f}}_i'^{(l)}}{\tau}\right)}\end{cases}\tag{7}$$

式中：τ 为温度系数，$L_{\text{in}}^{(l)}$ 为层内实例均值。

为使对比监督覆盖不同语义深度，采用归一化权重对各层损失进行加权求和：

$$\begin{cases}L_{\text{in}}=\sum\limits_{l=1}^{L_{\text{layer}}}\alpha_l L_{\text{in}}^{(l)}\\ \alpha_l\geqslant0,\ \sum\limits_{l}\alpha_l=1\end{cases}\tag{8}$$

式中：α_l 为第 l 层对比监督损失的归一化权重。

本文模型的权重 $\alpha_l\in[0,1]$ 用于平衡不同层级特征的语义一致性损失，为适配多视角BEV特征的层级差异，通过消融试验确定权重参数，对靠近输出层的高层特征赋予较高权重，以强调检测任务所需的强语义信息；同时采用指数加权方式，为浅层特征分配较低权重，以保留目标轮廓等细粒度细节。

2.4 透视区域对比学习

透视区域对比学习针对远距、小尺度与遮挡目标从图像域提升骨干网络局部表征质量。通过将BEV骨干网络标注投影至各视图，同时，在模型的中高层特征层对投影区域进行尺度自适应池化（Scale-Aware Pooling，SAP）操作，以增强不同尺度目标特征的鲁棒性，获取稳定的区域特征。在对比空间中，对不同视图或增强区域特征增加判别性约束。

将BEV空间中实例投影至第k个视图的像素框表示，并执行尺度自适应池化，则实例的池化特征为：

$$\boldsymbol{p}_{i,k} = SAP(\boldsymbol{P}_j^k, \boldsymbol{R}_{i,k}) = \frac{1}{Z_{i,k}} \sum_{u \in R_{i,k}} w_{i,k}(u) \boldsymbol{P}_j^k(u) \tag{9}$$

式中：$\boldsymbol{P}_j^k$为第k个视图的第j层特征图，$\boldsymbol{R}_{i,k}$为实例A_i在第k个视图上的投影像素框，$w_{i,k}(u)$为像素u的自适应权重，$Z_{i,k}$为权重归一化因子。

权重函数可由像素到区域中心距离的高斯函数结合像素面积归一化确定，归一化因子$Z_{i,k} = \sum_{u \in R_{i,k}} w_{i,k}(u)$。当投影面积极小时，可先采用插值放大或将该实例标注为难样本并在训练采样中降低权重以避免噪声污染。因此，池化后的区域特征经2层多层感知机$g(\cdot)$映射并归一化，得到特征$\tilde{\boldsymbol{p}}_{i,k} = g(\boldsymbol{p}_{i,k}) / \| g(\boldsymbol{p}_{i,k}) \|_2$后，在视图或增强对间构建正、负样本并计算对比损失：

$$\begin{cases} L_{\text{pers}} = \dfrac{1}{n} \sum_{i=1}^{n} l_{\text{pers},i} \\ l_{\text{pers},i} = -\dfrac{1}{|S_i|} \displaystyle\sum_{(k,k') \in S_i} \log \frac{\exp\left(\dfrac{\tilde{\boldsymbol{p}}_{i,k}^{\mathrm{T}} \tilde{\boldsymbol{p}}'_{i,k'}}{\tau} \right)}{\displaystyle\sum_{(u,v) \in N_{i,k}} \exp(\tilde{\boldsymbol{p}}_{i,k}^{\mathrm{T}} \tilde{\boldsymbol{p}}'_{u,v} / \tau)} \end{cases} \tag{10}$$

式中：S_i为正样本对集合，$N_{i,k}$为负样本集合，L_{pers}为透视区域对比损失。

2.5 检测头与损失函数

检测头负责输出目标类别概率、3D框中心偏移$(\Delta x, \Delta y, \Delta z)$、尺寸$(l, w, h)$、朝向角$\theta$和速度向量。本文使用BEVFormer检测头，保证检测任务的有效性和稳定性。基于可变形注意力（Deformable Attention，DA）机制的骨干网络模块，采用DETR架构中的解码器模块，用于多视角特征的空间交互与目标查询。

为进一步增强多视角特征表达的一致性与判别性，引入实例级对比损失和透视区域对比损失，构建统一的多任务优化目标函数。整体损失函数L主要包括检测主损失L_{det}、实例级对比损失L_{in}和透视区域对比损失L_{pers}：

$$L = \lambda_1 L_{\text{det}} + \lambda_2 L_{\text{in}} + \lambda_3 L_{\text{pers}} \tag{11}$$

式中：λ_1、λ_2、λ_3为各损失项的权重系数。

其中，检测主损失用于监督模型的分类与三维框回归任务；实例级对比损失用于约束同一物理目标在不同视角下BEV特征的一致性，缓解特征漂移；透视区域对比损失用于引导图像骨干网络学习三维定位相关特征。

2.5.1 检测主损失

检测主损失用于监督BEVFormer检测头的分类与回归分支，检测主损失由三维边界框回归损失L_{reg}、目标分类损失L_{cls}和方向角损失L_{dir}构成：

$$L_{\text{det}} = L_{\text{cls}} + \alpha_{\text{reg}} L_{\text{reg}} + \alpha_{\text{dir}} L_{\text{dir}} \tag{12}$$

式中：α_{reg}、α_{dir}分别为回归损失和方向角损失的权重系数。

分类损失采用标准交叉熵损失实现类别概率监督；回归损失采用平滑L_1损失交并比（Intersection over Union，IoU）损失加权融合以同步优化三维框位置、尺寸参数与框重叠精度，方向角损失采用分类与回归混合策略，兼顾角度周期性约束与回归精度。

2.5.2 实例级对比损失

实例级对比损失用于约束BEV空间中同一目标实例在不同视角或不同增强条件下的特征一致性。实例级对比损失为：

$$L_{\text{in}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_i^{+} / \tau)}{\exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_i^{+} / \tau) + \displaystyle\sum_{j} \exp(\boldsymbol{f}_i \cdot \boldsymbol{f}_j^{-} / \tau)} \tag{13}$$

式中：N为参与对比计算的实例数量，$\boldsymbol{f}_i$为对齐操作提取得的第i个目标实例的BEV特征表示，$\boldsymbol{f}_i^{+}$、$\boldsymbol{f}_j^{-}$分别为正、负样本集合。

通过该损失约束，可有效增强同一目标在BEV空间中的特征聚合性，抑制不同目标实例间的特征混淆。

2.5.3 透视区域对比损失

透视区域对比损失作用于多视角图像特征层，引导网络关注与三维目标检测任务相关的关键区域。透视区域对比损失为：

$$L_{\text{pers}} = -\frac{1}{M} \sum_{k=1}^{M} \log \frac{\exp(\boldsymbol{p}_k \cdot \boldsymbol{p}_k^{+} / \tau)}{\exp(\boldsymbol{p}_k \cdot \boldsymbol{p}_k^{+} / \tau) + \displaystyle\sum_{m} \exp(\boldsymbol{p}_k \cdot \boldsymbol{p}_m^{-} / \tau)} \tag{14}$$

式中：M为参与区域对比的样本数量，$\boldsymbol{p}_k$为第k个目标在图像空间中的透视投影区域特征，$\boldsymbol{p}_k^{+}$、$\boldsymbol{p}_m^{-}$分别为正、负样本区域特征集合。

该损失通过在图像特征层面引入区域级一致性约束，有助于提升远距离及小尺度目标在多视角条件下的特征稳定性，为后续BEV网络表征与检测任务提供输入特征。

3 试验环境与评估

3.1 数据集

采用nuScenes数据集对本文方法进行系统评估。该数据集包含1 000个城市真实驾驶场景，覆盖白天、夜晚、晴天、雨、雾等多种复杂交通条件。同时，数据集提供6个环视摄像头的多视角图像信息，并配套高精度

三维目标标注，涵盖车辆、行人、自行车等10类常见交通参与者。

本文严格遵循nuScenes官方划分规则，试验样本共85 201帧，按照7∶3划分训练集与验证集。

3.2 试验环境

本文试验使用图像处理器为英伟达GeForce RTX 3090，显存为24 GB，操作系统为Ubuntu 20.04 LTS，在深度学习框架PyTorch 2.8下运行环境。

选择ResNet-101[16]为图像特征提取骨干，使用ImageNet数据集预训练权重并进行初始化。多视角摄像机图像在预处理阶段采用缩放模式，输入分辨率与nuScenes数据集[17]保持一致。

BEV空间覆盖感知范围为(±50 m)×(±50 m)，网格分辨率为0.5 m，对应BEV特征尺寸为200像素×200像素，通道维度为256。模型训练采用带权重衰减的自适应矩估计AdamW优化器，初始学习率为2.0×10^{-4}，权重衰减系数为1.0×10^{-2}，总训练周期为24个。

3.3 评价指标

针对模型性能评估，将平均精度(mean Average Precision，mAP)A_{mAP}和综合检测得分指标(nuScenes Detection Score，NDS)A_{NDS}作为主要精度指标：

$$A_{\mathrm{NDS}}=\frac{1}{2}A_{\mathrm{mAP}}+\frac{1}{2}\left(1-\frac{A_{\mathrm{mATE}}+A_{\mathrm{mASE}}+A_{\mathrm{mAOE}}+A_{\mathrm{mAVE}}+A_{\mathrm{mAAE}}}{5}\right) \tag{15}$$

式中：A_{mATE}、A_{mASE}、A_{mAOE}、A_{mAVE}、A_{mAAE}分别为平均定位误差、平均尺寸误差、平均朝向误差、平均速度误差和平均属性误差。

其中，mAP衡量模型在不同目标类别的检测能力，其匹配准则基于三维目标中心距离，更符合自动驾驶场景对空间定位精度的实际需求；NDS通过对多项误差指标进行加权融合，对模型整体性能进行统一评价。

各误差指标均按照nuScenes官方实现进行归一化处理，使指标能够在保证检测精度评价的同时，综合反映目标定位、尺度估计及运动状态预测等关键性能。

4 试验结果与分析

4.1 对比试验

为验证本文方法的有效性，使用nuScenes验证集，将本文方法与BEVDet、FCOS3D、PGD、DETR3D、BEVFormer、Clip-bevformer[18]等主流多视角三维目标检测算法进行比较分析，其中，BEVDet为基于鸟瞰视角的单阶段检测方法，FCOS3D为全卷积单阶段三维检测框架，PGD采用概率化深度建模思路，DETR3D基于三维查询实现跨视图特征融合，Clip-bevformer在BEVFormer基础上引入真值流优化，结果如表1所示。

由表1可知，从整体指标看，本文方法在NDS和mAP上均取得了当前最优结果。相较于BEVFormer基线模型，本文方法的NDS、mAP分别提升0.4百分点和1.6百分点，说明在不改变原有BEVFormer时序建模与推理框架的前提下，引入面向表征学习的辅助监督机制，能够进一步增强BEV特征的判别性与检测精度。本文方法在各项指标上均有提升，表明该方法在多视角图像下具备良好的整体检测性能与鲁棒性。

进一步分析误差指标可知，与BEVFormer相比，本文方法的平均定位误差mATE降低至0.662，平均绝对误差下降8.7%，模型在三维目标空间位置估计的精度显著提升。该结果与本文的BEV编码器输出层引入实例级对比监督的设计目标高度一致，通过显式约束同一物理目标在不同视角与不同增强条件下的BEV表征一致性，有效缓解了多视角投影与视图变换过程中产生的实例级特征漂移问题。

本文方法在mAAE指标上也取得了一定改善，说明增强后的BEV表征有助于提升目标属性预测的稳定性。尽管本文在训练阶段引入多个辅助监督分支，但因模块在推理阶段均被移除，模型在保持BEVFormer原有推理延迟和显存占用时，能够实现感知精度的提升。

与同样引入表征增强策略的先进方法Clip-bevformer相比，本文方法的NDS、mAP分别提升0.5百

表1 不同三维目标检测方法在nuScenes验证集上的性能对比

方法	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
BEVDet	0.396	0.330	0.702	0.272	0.534	0.932	0.251
FCOS3D	0.415	0.343	0.725	0.263	0.422	1.292	0.153
PGD	0.428	0.369	0.683	0.260	0.439	1.268	0.185
DETR3D	0.425	0.346	0.773	0.268	0.383	0.842	0.216
BEVFormer	0.448	0.375	0.725	0.272	0.391	0.802	0.200
Clip-bevformer	0.447	0.385	0.688	0.264	0.393	0.862	0.185
本文	0.452	0.391	0.662	0.262	0.401	0.918	0.178

分点和0.6百分点，平均定位误差mATE降低0.026。模型性能差异主要源于：Clip-bevformer通过引入真实值流增强多视角特征的时空对齐能力，侧重从运动信息关联角度优化BEV表征，未涉及对比学习相关的特征判别性约束；而本文方法以对比学习为核心，构建图像域+BEV域双层级对比监督体系，以对比约束显式对齐跨视图目标特征，有效缓解了因视图转换导致的表征漂移问题，在检测精度与空间定位准确性上均可实现更优表现。

4.2 消融试验

为检测本文透视区域辅助监督与实例级一致性约束对系统的贡献，在nuScenes数据集的验证集上进行系统的消融试验。以BEVFormer为基线模型，选取反映空间感知质量的核心指标mAP、NDS、mATE、mASE和mAOE评估各模块性能，结果如表2所示。

表2 不同辅助监督模块对感知性能影响

编号	透视区域监督	实例级一致性	NDS	mAP	mATE	mASE	mAOE
1			0.448	0.375	0.725	0.272	0.391
2	√		0.451	0.380	0.707	0.268	0.395
3		√	0.450	0.378	0.684	0.270	0.398
4	√	√	0.452	0.391	0.662	0.262	0.401

注：√为使用该模块。

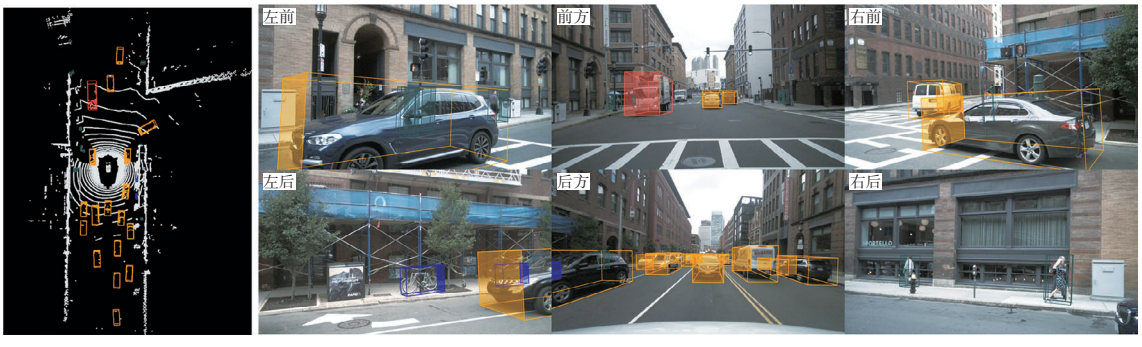
由表3分析可知，本文的2项机制在提升感知精度与优化空间定位方面体现出显著的互补性与协同效应。透视区域辅助监督通过在前端骨干网络引入几何约束，强化了模型对三维定位相关细粒度特征的提取能力，使mAP提升了0.5百分点。实例级一致性约束通过在中端BEV层显式对齐跨视角特征，有效缓解了视图变换引发的实例级几何漂移，使mATE显著降低0.041 m。

因此，本文方法能够实现从特征质量增强到空间对齐优化的全流程闭环，使mAP增益达1.6百分点，mATE进一步优化至0.662 m。此外，mASE的改善证实了精准特征对齐对物理边界推断的积极作用，而mAOE的平稳表现则反映出朝向预测对几何一致性约束的边际效益相对受限。

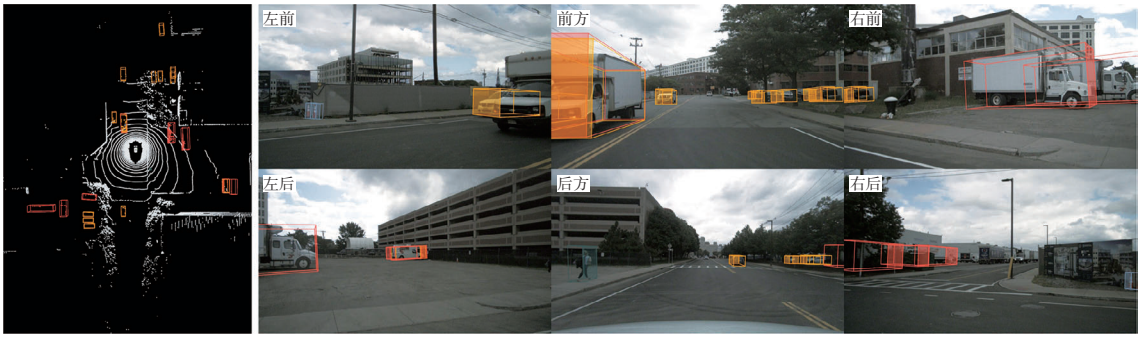
4.3 多视角图像目标检测可视化

为了直观展示本文方法在复杂交通场景下的感知性能，对nuScenes验证集中3种典型挑战场景的检测结果可视化，如图2所示。

由图2可知，在图2a中，由于引入实例级一致性约束，本文算法可有效缓解视图变换过程中的特征漂移，确保高密度目标在相机重叠区域的精准对齐。在图2b中，透视区域辅助监督机制引导骨干网络捕捉到更具判别性的局部细粒度特征，显著提升了模型对远端小目标的感知灵敏度。在图2c中，该机制进一步增强了模型在复杂光影干扰下的深度感知表征能力，使其仍能维持精确的空间定位与尺度估计。因此，可视化结果充分验证了本文算法在应对视图表征不一致及极端环境干扰方面的有效性与鲁棒性。



(a)城市峡谷场景



(b)阴、雾天场景



（c）夜晚场景

图2　不同场景下的BEV空间全局投影结果

对典型场景进行量化性能分析，如表3所示，本文方法在城市峡谷、阴雨天及夜晚场景下的mAP较BEVFormer均提升约2百分点，遮挡、远距与低照度关键目标的定位误差降低7%~8%，说明本文方法能够有效提升复杂驾驶场景下的检测精度与定位稳定性。

表3　典型场景量化性能对比

场景类型	mAP		mATE	
	BEVFormer	本文	BEVFormer	本文
城市峡谷	0.395	0.412	0.710	0.651
阴雨天	0.360	0.385	0.732	0.673
夜晚	0.348	0.370	0.740	0.685

5　结束语

本文以BEVFormer为基础，针对多视角三维目标检测中特征判别性弱、跨视图表征不一致的问题，构建了包含透视区域级辅助监督与实例级一致性约束的双层监督增强框架，从图像特征层与BEV表征层同步优化模型表达能力。该方法有效提升了检测精度与空间定位稳定性，对城市峡谷、恶劣天气、低照度等复杂驾驶场景的感知鲁棒性改善显著，可为自动驾驶多摄像头视觉感知提供有效技术参考。

后续研究可进一步结合更丰富的时空信息建模方式，探索轻量化表征增强策略，以持续提升多视角图像三维目标检测算法的鲁棒性与工程适用性。

参　考　文　献

[1] 黄德启，黄海峰，黄德意，等. BEV感知学习在自动驾驶中的应用综述[J]. 计算机工程与应用，2025，61(6)：1-21.
HUANG Q D，HUANG H F，HUANG D Y，et al. Review of Application of BEV Perceptual Learning in Autonomous Driving[J]. Computer Engineering and Applications，2025，61(6)：1-21.

[2] LIU Y F，WANG T C，ZHANG X X，et al. PETR：Position Embedding Transformation for Multi-View 3D Object Detection[C]// European Conference on Computer Vision. Aviv，Israel：Springer，2022：531-548.

[3] CAO H L，BEHNKE S. SLCF-Net：Sequential LiDAR-Camera Fusion for Semantic Scene Completion Using a 3D Recurrent U-Net[C]// 2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama，Japan：IEEE，2024：2767-2773.

[4] 庞亚军，张继文，宁一凡. 自动驾驶中LiDAR与相机融合的目标检测技术研究进展[J]. 汽车技术，2026(2)：18-31.
PANG Y J，ZHANG J W，NING Y F. Research Progress of Object Detection Technology Based on LiDAR-Camera Fusion in Autonomous Driving[J]. Automobile Technology，2026(2)：18-31.

[5] LIU H Y，WANG B B，ZHAGN L，et al. BEVoxSeg：BEV-Voxel Representation for Fast and Accurate Camera-Based 3D Segmentation[C]// ICASSP 2024-2024 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). Seoul，Korea：IEEE，2024：3345-3349.

[6] HUANG J J，HUANG G，ZHU Z，et al. BEVDet：High-Performance Multi-Camera 3D Object Detection in Bird-Eye-View[EB/OL]. (2022-06-12) [2026-02-26]. https://arxiv.org/abs/2112.11790.

[7] WANG Y，GUIZILINI V C，ZHANG T Y，et al. DETR3D：3D Object Detection from Multi-view Images via 3D-to-2D Queries[C]// Proceedings of the 5th Conference on Robot Learning. Seoul，Korea：PMLR，2022：180-191.

[8] WANG T，ZHU X G，PANG J M，et al. Probabilistic and Geometric Depth：Detecting Objects in Perspective[C]// Proceedings of the 5th Conference on Robot Learning. Seoul，Korea：PMLR，2022：1475-1485.

[9] WANG T，ZH1U X G，PANG J M，et al. Fcos3d：Fully Convolutional One-Stage Monocular 3D Object Detection[C]// 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal，QC，Canada：IEEE，2021.

[10] LI Z Q，WANG W H，LI H Y，et al. BEVFormer：Learning Bird's-Eye-View Representation from Lidar-Camera via

Spatiotemporal Transformers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 47(3): 2020-2036.

[11] DONG P Y, KONG Z L, MENG X, et al. HotBEV: Hardware-Oriented Transformer-Based Multi-View 3D Detector for BEV Perception[J]. Advances in Neural Information Processing Systems, 2023, 36: 2824-2836.

[12] CHEN T, KORNBLITH S, NOROUZI M, et al. A Simple Framework for Contrastive Learning of Visual Representations[C]// International Conference on Machine Learning. PMLR, 2020: 1597-1607.

[13] KHOSLA P, TETERWAK P, WANG C, et al. Supervised Contrastive Learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673.

[14] XIE E Z, DING J, WANG W H, et al. DetCo: Unsupervised Contrastive Learning for Object Detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021: 8392-8401.

[15] YU C H, ZHOU Q, WANG Z B, et al. ICPC: Instance-Conditioned Prompting with Contrastive Learning for Semantic Segmentation[EB/OL]. (2023-08-14) [2026-02-26]. https://arxiv.org/abs/2308.07078.

[16] HE K M, ZHANG X Y, REN S Q, et al. Deep Residual Learning for Image Recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 770-778.

[17] CAESAR H, BANKITI V, LANG A H, et al. NuScenes: A Multimodal Dataset for Autonomous Driving[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 11621-11631.

[18] PAN C B, YAMAN B, VELIPASALAR S, et al. CLIP-BEVFormer: Enhancing Multi-View Image-Based Bev Detector with Ground Truth Flow[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2024: 15216-15225.

（责任编辑　瑞　秋）

修改稿收到日期为2026年2月26日。

《汽车技术》征稿启事

《汽车技术》杂志是中国第一汽车集团有限公司主办的国内外公开发行的汽车前瞻与应用技术类月刊，为我国高质量科技期刊分级目录入选期刊、中国科学引文数据库（CSCD）来源期刊、中文核心期刊、中国科技核心期刊、RCCSE中国核心学术期刊（A）、Scopus数据库收录期刊、俄罗斯《文摘杂志》（AJ）收录期刊、日本科学技术振兴机构数据库入选期刊、EBSCO学术数据库收录期刊、欧洲学术出版中心（EuroPub）数据库收录期刊。

《汽车技术》杂志以报道汽车整车及其零部件设计、研究、试验等方面的前瞻与应用技术为主，并兼有理论研究内容，是中国汽车行业核心学术和知识传播与共享的平台。

《汽车技术》将在国家提出的“创新、协调、绿色、开放、共享”发展理念的指引下，把握《节能与新能源汽车技术路线图》和“低碳化、信息化、智能化”的汽车技术主流发展趋势，努力在传统内燃机汽车高效动力系统、轻量化、低阻力领域，新能源汽车和互联智能汽车技术领域，大力吸收优质稿源，为广大科研和工程技术人员服务，为我国汽车工程技术创新能力提升贡献力量。

《汽车技术》欢迎高等院校师生、研发工程技术人员、技术管理人员及相关人员不吝赐稿，反映国家重点扶持项目、自然科学基金项目和其他重点项目等研究成果的稿件将被优先选择刊登。

投稿要求：

1. 文章字数最好控制在6 000~8 000字范围之内；

2. 请按科技论文要求撰写文章摘要，摘要中文字数控制在180字左右；

3. 文章必须附有公开发表的、体现本领域最新研究成果的参考文献，且在文中应标注文献引用处；

4. 文章主要作者应提供其简介，包括出生年、性别、职称、学历、研究方向及技术成果等；

5. 来稿的保密审查工作由作者单位负责，确保署名无争议，文责自负；

6. 请勿一稿多投；

7. 本刊使用网站投稿，请先登陆网站注册成功后投稿，详细投稿要求见本刊网站中“下载中心”栏的“作者指南”，网址：http://qcjs.cbpt.cnki.net。

《汽车技术》编辑部