计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(8): 2059−2075, 2026

# 基于动态缩放卷积的多模态 3D 目标检测

纪文宇　苗　壮　张启阳　崔小璐　李唯暄　李　阳
（中国人民解放军陆军工程大学指挥控制工程学院　南京　210007）
（wenyuji_edu@foxmail.com）

# Multi-Modal 3D Object Detection Based on Dynamic Scaling Convolution

Ji Wenyu, Miao Zhuang, Zhang Qiyang, Cui Xiaolu, Li Weixuan, and Li Yang
（*College of Command and Control Engineering*, *Army Engineering University of PLA*, *Nanjing* 210007）

**Abstract**　Multi-modal 3D object detection is a pivotal task in autonomous driving and environment perception. It achieves accurate localization and recognition of objects in 3D space by integrating data from multiple sensors, such as LiDAR (light detection and ranging) point clouds and camera images. To address the common challenges of large scale variations and insufficient multi-scale feature extraction in existing methods, we propose a multi-modal 3D object detection method based on dynamic scaling convolution. For this method, we design a dual-branch multi-modal fusion architecture, which consists of a VoxelNet-based 3D backbone and a novel dynamic scaling network. The two branches collaboratively process LiDAR point clouds and image-derived virtual point clouds to achieve cross-modal feature alignment and adaptive fusion. The core of the dynamic scaling network is a dynamic scaling module, which integrates 3D and 2D dynamic scaling convolution operators specifically tailored for multi-modal feature processing. This module adaptively adjusts the sampling locations and the number of valid sampling points within the receptive fields of 3D and 2D branches, thus effectively capturing structural and semantic variations of multi-scale objects. Extensive experiments on two benchmark datasets demonstrate the proposed method outperforms more than ten prevailing methods in terms of AP (average precision). Specifically, it achieves $mAP_{3D}$ (3D mean average precision) scores of 86.87%, 46.68%, and 68.39% for the car, pedestrian, and cyclist categories on KITTI test set, and achieves a mAP (mean average precision) of 71.8% on nuScenes test set. These results demonstrate the superior detection performance and strong generalization ability of the proposed method for multi-modal 3D object detection tasks.

**Key words**　autonomous driving; LiDAR point cloud; camera image; multi-modal fusion; 3D object detection

**摘　要**　作为自动驾驶与环境感知中的核心任务，多模态 3D 目标检测通过融合激光雷达点云与相机图像等多种传感器数据，来实现对 3D 空间目标的精准定位与识别。为解决现有方法中普遍存在的目标规模差异与多尺度特征提取不足问题，提出了一种基于动态缩放卷积的多模态 3D 目标检测方法。该方法设计了一种多模态融合双分支模型，由体素网络和所设计的动态缩放网络组成，能够协同处理激光雷达点云和图像衍生的虚拟点云，实现跨模态特征对齐与自适应融合。动态缩放网络的核心是动态缩放模块，该模块包含适用于多模态数据的 3D 动态缩放卷积和 2D 动态缩放卷积，能够在 3D 与 2D 分支上自适应地调整采样位置和感受野内有效采样点数量，捕捉多尺度目标的结构与语义差异。在 2 个数据集上与现有 10 余种主流方法的对比结果表明，该方法的平均精度（AP）均超过现有方法。具体地，在 KITTI 测试集汽车、行人和骑行者类别上分别达到 86.87%, 46.68%, 68.39% 的 3D 目标检测平均精度均值（$mAP_{3D}$），在 nuScenes 测试集上达到 71.8% 的平均精度均值（mAP），有效证明了所提方法在多模态 3D 目标检测任务

**收稿日期**：2025-09-21；**修回日期**：2026-04-13
**基金项目**：国家自然科学基金项目（62273356）
This work was supported by the National Natural Science Foundation of China (62273356).
**通信作者**：李阳（solarleeon@outlook.com）

中的检测与泛化性能。

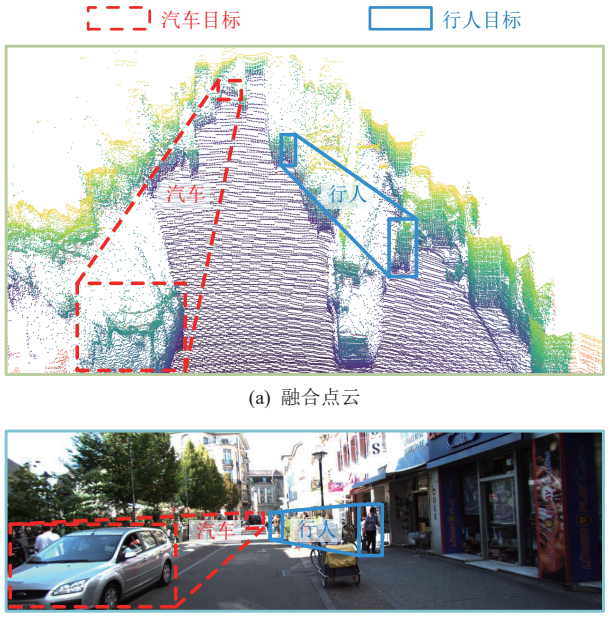
**关键词** 自动驾驶;激光雷达点云;相机图像;多模态融合;3D 目标检测

**中图法分类号** TP391.41

**DOI:** 10.7544/issn1000-1239.202550676 **CSTR:** 32373.14.issn1000-1239.202550676

精准的 3D 目标检测在自动驾驶、机器人导航、增强现实等多个领域中至关重要[1-4]。基于激光雷达(light detection and ranging, LiDAR)的 3D 目标检测方法虽能提供精确的深度信息[5-7],但在目标纹理、颜色识别以及恶劣天气或远距离场景下,难以保证足够的鲁棒性和精度。因此,一些方法开始尝试多模态 3D 目标检测[8-9],通过对 3D 激光雷达与 2D 相机进行融合,实现优势互补。多模态方法能够借助激光雷达的几何感知能力与相机图像的丰富纹理信息[10-12],有效提升复杂场景下 3D 目标的检测性能。

在多模态 3D 目标检测中,不同目标的尺度存在显著差异,例如汽车、骑行者和行人等[13]。即使是同种目标,由于与激光雷达、相机的距离不同,在点云或图像中呈现的尺度也有明显差异。如图 1 所示,对于远距离和近距离的汽车目标、行人目标,如果使用固定尺寸的标准卷积对点云或图像数据进行特征提取,将无法自适应地调整感受野以匹配不同目标的尺度特性[14-15]。这是因为标准卷积不仅在空间上具有固定的采样位置,而且卷积核大小和采样点数量通常是预先定义的、静态的,难以捕捉多尺度、多形态目标的精细化特征,导致对不同尺度目标的检测性能下降。




(a) 融合点云

(b) 相机图像

Fig. 1 Comparison of scale differences among different objects

图 1 不同目标的尺度差异对比

在标准卷积的基础上,尽管一些新颖的卷积方法通过引入自适应机制或扩大感受野来提升模型感知能力[16-17],能够在特定模态或任务中取得更高效或更精确的表现,但是这些方法并未充分考虑多模态数据处理面临的计算负担或复杂场景下不同目标的尺度差异。同时,这些方法通常在单一维度(如 2D 图像空间或 3D 点云空间)上进行操作[18-19],难以实现不同维度下异构特征交互或自适应融合的功能,无法应对多模态 3D 目标检测在特征表达一致性、对齐精度及计算效率方面的较高需求。

为了解决上述问题,受主动视觉感知[20-22]启发,本文提出了一种基于动态缩放卷积(dynamic scaling convolution, DSConv)的多模态 3D 目标检测方法。该方法通过体素网络与所设计的动态缩放网络构建了双分支网络架构,能够以互补方式处理激光雷达点云与图像衍生的虚拟点云,实现多模态特征融合并进行 3D 目标检测。本文提出的动态缩放网络由 4 层动态缩放模块级联构成,该模块的核心操作包含 3D 动态缩放卷积和 2D 动态缩放卷积。动态缩放卷积不仅能够根据输入的特征内容动态调整采样位置和感受野内有效采样点数量,实现多尺度目标特征的自适应捕获;而且考虑了不同模态数据的特点,能够兼顾 3D 激光雷达点云的稀疏性与无序性、2D 相机图像的密集性与局部相关性。实验结果表明,本文方法在 KITTI 测试集汽车、行人和骑行者类别上分别达到 86.87%, 46.68%, 68.39% 的 3D 目标检测平均精度均值(3D mean average precision, $mAP_{3D}$),在 nuScenes 测试集上达到 71.8% 的平均精度均值(mean average precision, mAP),均超过现有的 10 余种主流方法,验证了该方法捕捉不同目标尺度差异、提升多模态 3D 目标检测精度的有效性。

本文贡献主要包括 4 个方面:

1)提出了一种基于动态缩放卷积的多模态 3D 目标检测方法,该方法能够根据输入内容自适应地调整采样位置与感受野内有效采样点数量,显著提

升多尺度目标的特征捕获能力。

2)提出了一种双分支网络架构，通过体素网络与动态缩放网络协同设计，实现了激光雷达点云与图像虚拟点云在特征层面的互补性对齐与自适应融合。

3)提出了兼顾 3D 点云稀疏性与 2D 图像局部相关性的动态缩放卷积，通过级联式结构同时支持 3D 与 2D 卷积操作，有效克服了多模态数据的结构差异问题。

4)在权威的 KITTI 和 nuScenes 数据集上取得了具有竞争力的综合检测性能。实验结果表明，所提方法在多数关键指标上优于现有主流方法，整体性能表现突出。

## 1 相关工作

### 1.1 激光雷达 3D 目标检测

近年来，基于激光雷达的 3D 目标检测逐步从体素化建模向高效特征学习演进[23-26]。VoxelNet[14] 实现点云体素化并结合 3D 卷积进行端到端学习，为该领域奠定了基础。为提升定位精度，PV-RCNN[27] 融合体素与点特征进行细粒度建模，而 SA-SSD[28] 通过引入辅助网络与特征对齐机制，增强了结构感知能力。此外，Point-GNN[29] 通过图结构与自注册机制，解决了空间变换中点云数据的特征不稳定问题。然而，这些方法仍受限于点云稀疏性[30]，在小目标与遮挡场景中检测精度不足。因此，多模态 3D 目标检测成为突破瓶颈的重要方向。

### 1.2 多模态 3D 目标检测

多模态 3D 目标检测通过融合图像与点云信息提升感知鲁棒性[31-33]，其核心挑战在于跨模态特征对齐与高效交互。根据融合时机，现有方法可分为早期融合、中期融合与晚期融合。F-PointNet[34] 和 F-ConvNet[35] 等早期融合方法通过视锥体提取局部点云并与图像特征进行融合，但其几何投影仅能实现粗略对齐；AutoAlign[36]，VirpNet[37]，Is-fusion[38] 等中期融合方法在特征层面利用可学习对齐模块与交叉注意力进行自适应交互，实现了语义级对齐；Fast-CLOCs[39] 和 C-CLOCs[40] 等晚期融合方法则在检测结果层面进行融合，但不同模态差异容易导致融合偏差。在特征对齐机制上，现有工作主要分为几何驱动、特征学习与虚拟生成 3 类。EPNet[12]，PointPainting[41] 等几何驱动方法依赖传感器标定与投影，计算高效但受限于标定误差；UniTR[42]，SparseFusion[43] 等特征学习方法通过注意力或可学习模块隐式对齐，灵活性高但需要大量训练；相比之下，MVP[44]，VirConv[45] 等虚拟生成方法将图像语义映射为虚拟点云，有效补充了远距离目标的特征表达。本文综合考虑特征对齐效率与远距离目标感知能力，采用激光雷达点云与图像虚拟点云融合的方法，通过跨模态特征交互解决点云稀疏问题。

### 1.3 3D 卷积与 2D 卷积

3D 卷积与 2D 卷积作为深度学习中的核心操作，近年来在多模态感知领域取得了显著进展[46-50]。3D 卷积通过焦点稀疏卷积[51] 和大核扩展[52] 等方法优化了点云特征提取效率，但 Ada3D[16] 和 Relation3D[18] 等方法揭示其对多尺度目标的空间冗余处理仍存在局限。PVNAS[53] 虽然通过神经架构搜索平衡了模型复杂度与精度，但其采用的固定感受野难以适应不同目标尺度差异。2D 卷积通过动态注意力机制[15] 和可分离结构[54] 等方法降低了计算成本，LSKNet[17] 和 WTConv[19] 通过频域分析与选择性内核增强了长程建模能力。PConv[55] 利用非对称填充扩展感受野，RefConv[56] 通过重参数化提升小目标识别鲁棒性。然而，2D 卷积方法难以满足 3D 场景中几何信息的精确建模需求。为解决 3D 卷积的尺度敏感性与 2D 卷积的几何表达缺陷，本文提出了 3D 动态缩放卷积和 2D 动态缩放卷积。如图 2 所示，相较于标准卷积，动态缩放卷积通过自适应调整采样位置和感受野内有效采样点数量，能够有效捕捉多模态 3D 目标检测中不同目标的尺度差异。

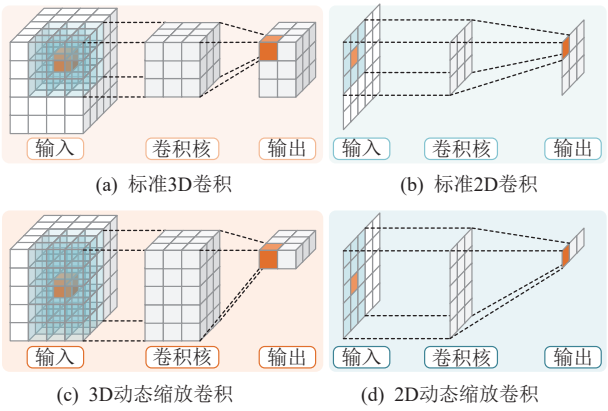



Fig. 2 Comparison between standard convolution and dynamic scaling convolution

图 2 标准卷积与动态缩放卷积对比

## 2 本文方法

### 2.1 双分支网络架构

为解决不同目标尺度差异显著、特征提取效果

不佳问题，本文受 VirConv-T 模型[45]的启发，利用包含 3D 动态缩放卷积和 2D 动态缩放卷积的动态缩放模块，构建了动态缩放网络，进而与体素网络 VoxelNet[14]组合形成了双分支网络架构，如图 3 所示。与细化方法 CasA[23]以及多重变换设计方法 TED[57]不同，我们构建的双分支网络架构仍采用后期融合方法以及变换精炼方案，以降低计算负担。

具体来说，我们通过图像引导的深度补全方法生成虚拟点云。遵循以往的工作[45,58]，我们采用 PENet[59]从相机图像和稀疏激光雷达点云中预测密集深度图，并利用 MVP[44]的多模态虚拟点生成策略，将预测的深度信息反投影到 3D 空间，从而获得图像虚拟点云。在对激光雷达点云和图像虚拟点云进行融合后，可以得到包含更丰富几何信息的融合点云。对于融合点云，我们利用 2 个不同的分支分别进行处理，其中第 1 分支使用旋转变换，第 2 分支使用反射变换。其次，分别使用体素网络和我们所设计的动态缩放网络对融合点云特征进行编码。为了降低计算成本并提升特征学习一致性，我们在 2 个分支之间共享卷积权重。然后，借助区域提案网络(RPN)生成感兴趣区域，并利用旋转变换下的体素网络特征和动态缩放网络特征进行精炼，进而利用反射变换下的体素网络特征和动态缩放网络特征进一步增益精炼后的感兴趣区域。最后，我们利用边界框投票法对精炼后的感兴趣区域进行融合，并利用非极大值抑制(NMS)得到检测结果。

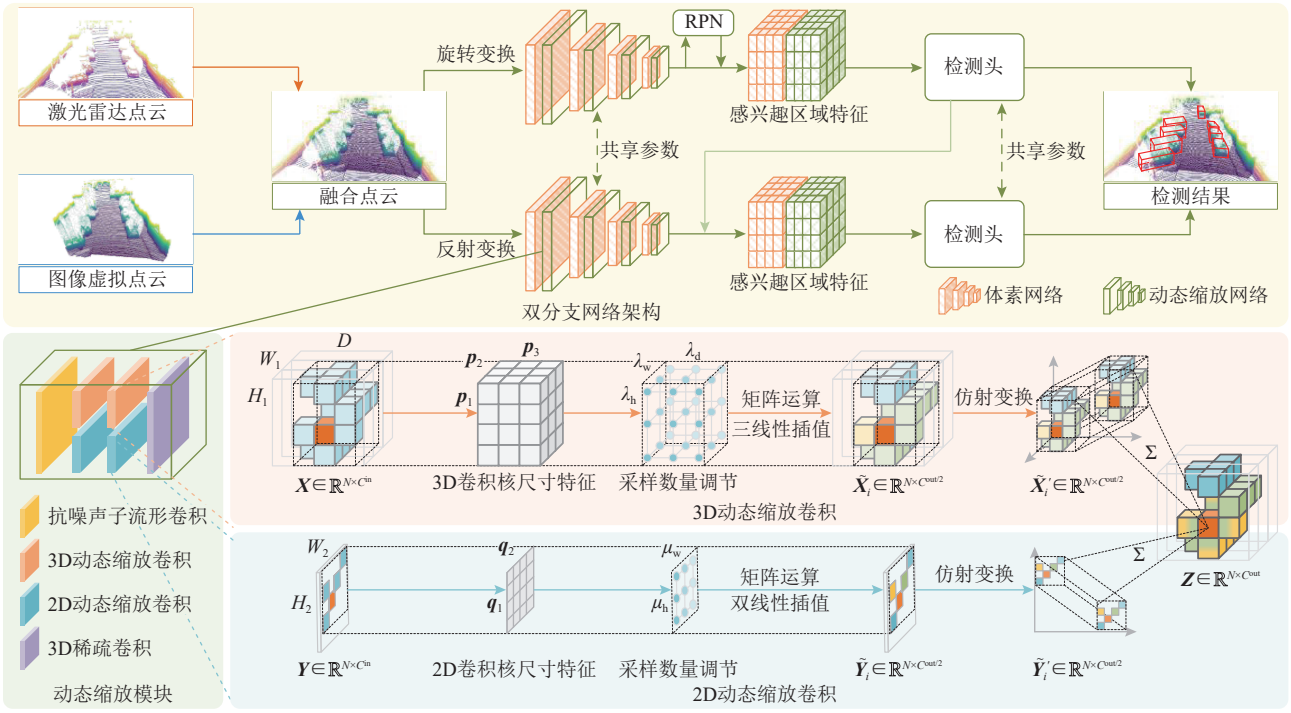



Fig. 3 Network framework of the proposed method

图 3 本文方法的网络框架

## 2.2 动态缩放模块

如图 3 所示，在双分支网络架构中，动态缩放网络的核心是 4 个下采样步长分别为 1, 2, 4, 8 的动态缩放模块。动态缩放模块由 4 个部分组成：1)1 层抗噪声子流形卷积[45]，用于编码不同维度特征并滤除虚拟点噪声；2)2 层 3D 动态缩放卷积，用于自适应提取激光雷达点云和图像虚拟点云中不同目标的尺度特征；3)2 层 2D 动态缩放卷积，用于自适应提取图像中不同目标的尺度特征；4)1 层 3D 稀疏卷积[60-61]，用于对特征结果进行下采样。

当融合点云输入动态缩放模块后，首先会经过抗噪声子流形卷积层。该卷积层能够利用 3D 空间和 2D 图像滤除不准确深度估计带来的噪声[45,62]。记高度 $H_1$、宽度 $W_1$ 以及深度 $D$ 构成的 3D 体素空间中，非空体素的总数为 $N$，输入抗噪声子流形卷积的含噪特征为 $\boldsymbol{X}_{\text{noise}} \in \mathbb{R}^{N\times C^{\text{in}}}$。其中，$N$ 满足 $N \leqslant H_1 \times W_1 \times D$，$C^{\text{in}}$ 表示输入特征通道的数量。经过抗噪声子流形卷积处理后，输出的 3D 抗噪特征通道数不会发生改变，因此可以表示为 $\boldsymbol{X} \in \mathbb{R}^{N\times C^{\text{in}}}$：

$$\boldsymbol{X} = filter_{\text{3D,2D}}(\boldsymbol{X}_{\text{noise}}), \tag{1}$$

其中$filter_{3D,2D}(\cdot)$表示抗噪声子流形卷积操作。

为了有效增强 3D 卷积核对不同目标尺度差异的捕捉能力，在得到 3D 抗噪特征$\boldsymbol{X}$后，利用 3D 动态缩放卷积对 3D 卷积核的采样位置和采样点数量进行动态缩放，获得 3D 采样特征后进行仿射变换，以增强动态缩放的灵活性和鲁棒性。最终，得到处理后的 3D 尺度特征$\tilde{\boldsymbol{X}}_i' \in \mathbb{R}^{N\times C^{\text{out}/2}}$：

$$\tilde{\boldsymbol{X}}_i' = DSConv_{3D}(\boldsymbol{X}), \tag{2}$$

其中$DSConv_{3D}(\cdot)$表示 3D 动态缩放卷积操作。

在抗噪过程中，抗噪声子流形卷积能够区分 2D 图像实例边界上的噪声，进而对噪声进行了有效抑制[45]。然而，抗噪声子流形卷积引入了 2D 图像的目标尺度差异问题。为解决该问题，对 3D 动态缩放卷积进行降维，获得 2D 动态缩放卷积。记输入 2D 动态缩放卷积的 2D 抗噪特征为$\boldsymbol{Y} \in \mathbb{R}^{N\times C^{\text{in}}}$，同理可得 2D 尺度特征$\tilde{\boldsymbol{Y}}_i' \in \mathbb{R}^{N\times C^{\text{out}/2}}$：

$$\tilde{\boldsymbol{Y}}_i' = DSConv_{2D}(\boldsymbol{Y}), \tag{3}$$

其中$DSConv_{2D}(\cdot)$表示 2D 动态缩放卷积操作。

获得 3D 尺度特征和 2D 尺度特征后，本文采用拼接的方式进行处理，以同时应对 3D 和 2D 维度上的目标尺度差异问题：

$$\boldsymbol{Z} = [\tilde{\boldsymbol{X}}_i', \tilde{\boldsymbol{Y}}_i'], \tag{4}$$

其中$\boldsymbol{Z} \in \mathbb{R}^{N\times C^{\text{out}}}$表示拼接后的特征。

为了减少对空区域的计算，降低内存和计算资源消耗，本文利用了 SparseConv3D[60] 和 SubMConv3D[61] 这 2 种 3D 稀疏卷积，可以在保持稀疏性的同时扩大感受野。最后，本文将拼接后的特征$\boldsymbol{Z}$输入到 3D 稀疏卷积中进行下采样：

$$\boldsymbol{Z}' = SpConv_{3D}(\boldsymbol{Z}), \tag{5}$$

其中$SpConv_{3D}(\cdot)$表示 3D 稀疏卷积操作，$\boldsymbol{Z}'$表示下采样后的特征。

由于 3D 动态缩放卷积和 2D 动态缩放卷积是动态缩放模块的关键结构，因此本文将在 2.3 节介绍 3D 动态缩放卷积，在 2.4 节介绍 2D 动态缩放卷积。

## 2.3 3D 动态缩放卷积

为应对 3D 目标在点云场景下的尺度差异，我们设计了 3D 动态缩放卷积。3D 动态缩放卷积针对抗噪声子流形卷积层输出的 3D 抗噪特征，首先动态地改变卷积核尺寸，调节采样点数量；其次，利用矩阵运算和三线性插值生成 3D 采样特征；然后对 3D 特征进行仿射变换，增强动态缩放卷积在点云场景下的空间适应性；最后对尺度特征进行拼接，获取动态缩放特征并传入 3D 稀疏卷积层。下面，本文对 3D 动态缩放卷积的细节部分进行逐步阐述。

### 2.3.1 3D 卷积核参数调节

受 ARConv[63] 的启发，本文设计了用于调节 3D 卷积核尺寸的 3D 动态缩放器$f_i(\cdot)$，如图 4 所示。将 3D 抗噪特征$\boldsymbol{X}$输入到 3D 动态缩放器$f_i(\cdot)$后，能够获取调节后 3D 卷积核的尺寸特征$\boldsymbol{p}_i \in \mathbb{R}^{N\times 1}$：

$$\boldsymbol{p}_i = f_i(\boldsymbol{X}),\ i \in \{1,2,3\}, \tag{6}$$

其中$\boldsymbol{p}_1$表示高度特征，$\boldsymbol{p}_2$表示宽度特征，$\boldsymbol{p}_3$表示深度特征。

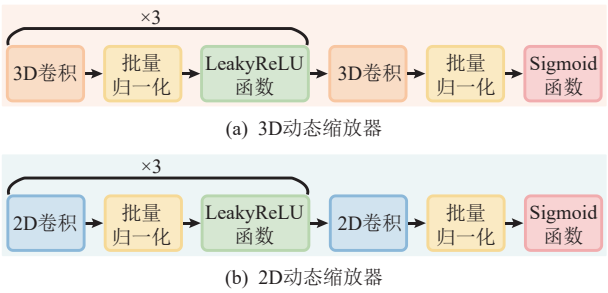



Fig. 4 Dynamic scalers

图 4 动态缩放器

在 3D 动态缩放器的最后 1 层，利用 Sigmoid 函数控制高度、宽度和深度的缩放程度，确保数值为正且有界。计算$\boldsymbol{p}_1$，$\boldsymbol{p}_2$，$\boldsymbol{p}_3$的平均值$\bar{p}_1$，$\bar{p}_2$，$\bar{p}_3$，获取 3D 卷积核的平均尺寸。记 3D 卷积核在高度、宽度和深度上的缩放系数分别为$\alpha_1$，$\beta_1$，$\gamma$，那么初步调节后的采样点数为$\frac{\bar{p}_1}{\alpha_1}$，$\frac{\bar{p}_2}{\beta_1}$，$\frac{\bar{p}_3}{\gamma}$。当$\alpha_1$，$\beta_1$，$\gamma$的数值越大时，缩放后采样点数量就越少，此时分布呈现稀疏状态；当$\alpha_1$，$\beta_1$，$\gamma$的数值越小时，缩放后采样点数量就越多，此时分布呈现密集状态。考虑到采样点数可能为非整数或偶数，因此使用函数$\phi(\cdot)$进行修正，得到不大于当前数值的最大奇数：

$$\phi(x) = 2\times\left\lfloor\frac{\lfloor x\rfloor - 1}{2}\right\rfloor + 1, \tag{7}$$

其中“$\lfloor\cdot\rfloor$”表示向下取整函数。

修正后高度、宽度和深度维度上的采样点数量分别为$\lambda_{\text{h}} = \phi\left(\frac{\bar{p}_1}{\alpha_1}\right)$，$\lambda_{\text{w}} = \phi\left(\frac{\bar{p}_2}{\beta_1}\right)$，$\lambda_{\text{d}} = \phi\left(\frac{\bar{p}_3}{\gamma}\right)$，均为正奇数。与 2D 卷积核一致，3D 卷积核的尺寸也倾向于设置为奇数，这是因为奇数尺寸可以明确几何中心、降低参数量并简化对称填充操作。据此，总采样点数量为

$$S_1 = \lambda_{\text{h}}\times\lambda_{\text{w}}\times\lambda_{\text{d}}。 \tag{8}$$

### 2.3.2 3D 矩阵运算及插值

在标准 3D 卷积中，采样点位置围绕卷积核中心

呈规则网格分布,因此可用偏移矩阵$\boldsymbol{U}$来计算采样点偏离3D卷积核中心的位置。记偏移矩阵$\boldsymbol{U}$中第$i$行、第$j$列、第$k$层(对应3D索引$(i,j,k)$)的元素为

$$u_{ijk}=\left(i-\frac{\lambda_{\mathrm{h}}+1}{2},j-\frac{\lambda_{\mathrm{w}}+1}{2},k-\frac{\lambda_{\mathrm{d}}+1}{2}\right)。\tag{9}$$

假设点云空间中某一位置为$\boldsymbol{p}_0$,动态缩放卷积核在$\boldsymbol{p}_0$处的高度为$h_0$、宽度为$w_0$、深度为$d_0$,根据上述采样参数$\lambda_{\mathrm{h}}$, $\lambda_{\mathrm{w}}$, $\lambda_{\mathrm{d}}$,可以得到比例矩阵$\boldsymbol{X}_0$。进一步,记比例矩阵$\boldsymbol{X}_0$中第$i$行、第$j$列、第$k$层的元素为

$$\boldsymbol{x}_{ijk}=\left(\frac{h_0}{\lambda_{\mathrm{h}}},\frac{w_0}{\lambda_{\mathrm{w}}},\frac{d_0}{\lambda_{\mathrm{d}}}\right)。\tag{10}$$

为了更灵活地捕捉融合点云特征,突破固定采样模式的限制,可以对偏移矩阵和比例矩阵进行运算,获取3D动态缩放卷积的动态采样矩阵:

$$\hat{\boldsymbol{R}}=\boldsymbol{U}\odot\boldsymbol{X}_0,\tag{11}$$

其中"$\odot$"表示逐元素相乘。

在体素空间中,目标点的坐标$(x,y,z)$通常不与网格点重合,需通过周围8个网格点插值得到目标点值。因此,我们参照以往的工作[64],使用三线性插值[65]处理3D空间中的非网格点采样,以保持数据的连续性和精度。而后,3D动态缩放卷积运算可以表示为

$$\Psi(\boldsymbol{p}_0)=\sum_{\hat{r}\in\hat{\boldsymbol{R}}}\sigma(\hat{r})\times\hat{t}(\boldsymbol{p}_0+\hat{r}),\tag{12}$$

其中$\Psi(\boldsymbol{p}_0)$表示$\boldsymbol{p}_0$处输出的3D特征,$\sigma$表示3D卷积核参数,$\hat{r}$表示动态采样矩阵$\hat{\boldsymbol{R}}$中的元素,$\hat{t}(\boldsymbol{p}_0+\hat{r})$用于计算$(\boldsymbol{p}_0+\hat{r})$处的采样值。

#### 2.3.3 3D采样特征生成及仿射变换

由于已经实现位置$\boldsymbol{p}_0$处的3D动态缩放卷积运算,那么对于所有采样点$S_1$中的每个位置进行计算,即可得到新的3D网格$\boldsymbol{P}_0$。进一步地,对体素空间中的所有位置重复上述运算,生成3D采样特征$\tilde{\boldsymbol{X}}_i\in\mathbb{R}^{N\times C^{\mathrm{out}/2}}$:

$$\tilde{\boldsymbol{X}}_i=\Psi(\boldsymbol{P}_0)=\sum\Psi(\boldsymbol{p}_0),\tag{13}$$

其中$C^{\mathrm{out}}$表示输出特征通道的数量,$C^{\mathrm{out}/2}$表示输出特征通道数经减半后得到的通道数量。

为对齐不同位置特征、增强特征表达多样性,让模型更好地捕捉3D数据的空间与通道关联,我们对3D采样特征$\tilde{\boldsymbol{X}}_i$进行仿射变换,仿射变换公式为:

$$\tilde{\boldsymbol{X}}_i'=\boldsymbol{A}_1\tilde{\boldsymbol{X}}_i+\boldsymbol{B}_1,\tag{14}$$

其中$\tilde{\boldsymbol{X}}_i'\in\mathbb{R}^{N\times C^{\mathrm{out}/2}}$表示经仿射变换后的3D尺度特征,$\boldsymbol{A}_1$表示仿射变换的线性变换矩阵,$\boldsymbol{B}_1$表示仿射变换的平移向量。

### 2.4 2D动态缩放卷积

为了滤除不准确深度估计导致的噪声,抗噪声子流形卷积层[45]引入了3D点云场景对应的2D图像,这带来了2D图像中的目标尺度差异问题。针对上述问题,我们对3D动态缩放卷积进行降维,得到适合处理图像的2D动态缩放卷积。具体来说,2D动态缩放卷积针对抗噪声子流形卷积层输出的2D抗噪特征,首先动态地改变卷积核的尺寸,调节采样点的数量;其次,利用矩阵运算和双线性插值[66]生成2D采样特征;然后对2D特征进行仿射变换,增强动态缩放卷积在2D图像中的适应性;最后进行拼接,获取动态缩放后的特征结果并传入3D稀疏卷积层。

具体而言,对于激光雷达到相机投影变换得到的2D像素空间,其高度记为$H_2$、宽度记为$W_2$。由于单一非空体素中可能存在多个特征,我们对相关体素进行最大池化,进而保留唯一特征进行2D动态缩放卷积。如图4所示,为了便于2D卷积核参数调节,我们对部分结构进行了调整,从而把3D动态缩放器降维成2D动态缩放器。向2D动态缩放器中输入2D抗噪特征$\boldsymbol{Y}$后,能够获取调节后的2D卷积核高度特征$\boldsymbol{q}_1$和宽度特征$\boldsymbol{q}_2$,进而计算平均值$\bar{q}_1$和$\bar{q}_2$以获取2D卷积核的平均尺寸。3D动态缩放卷积的原理类似,本文对2D卷积核参数进行调节,进而利用式(7)中的函数$\phi(\cdot)$进行尺度修正,得到修正后高度和宽度维度上的采样点数量$\mu_{\mathrm{h}}$和$\mu_{\mathrm{w}}$,二者均为正奇数。

与标准3D卷积一致,标准2D卷积采样点的位置也围绕卷积核中心呈规则网格分布。因此,本文对标准2D卷积展开矩阵运算,同时利用双线性插值[66]处理2D图像中的非网格点采样。对像素空间中的位置进行运算后,能够得到2D采样特征$\tilde{\boldsymbol{Y}}_i\in\mathbb{R}^{N\times C^{\mathrm{out}/2}}$。其中$C^{\mathrm{out}/2}$表示输出特征通道数经减半后得到的通道数量。为了对齐不同位置特征、增强特征表达多样性,让模型更好地捕捉2D数据的空间与通道关联,我们参照3D动态缩放卷积操作,对2D采样特征进行仿射变换,得到仿射变换后的2D尺度特征$\tilde{\boldsymbol{Y}}_i'$。最终,利用式(4)拼接3D尺度特征与2D尺度特征,而后传入3D稀疏卷积层进行特征下采样。

## 3 实 验

### 3.1 数据集和评价指标

#### 3.1.1 数据集

本文在KITTI[67]和nuScenes[68]数据集上对所提方法进行了全面评估。KITTI数据集是最早用于自动驾驶的3D目标检测数据集之一,其包含7 481个训练样本和7 518个测试样本,每个样本配备相机图

像、激光雷达点云和标定矩阵。为了评估本文方法的性能，我们遵循以往的策略[45]，将训练样本划分为包含 3 712 个样本的训练集以及包含 3 769 个样本的验证集。此外，KITTI 数据集根据目标的形状、遮挡级别和距离等因素，将所有目标分为简单、中等和困难 3 类。nuScenes 数据集是由 NVIDIA 支持的大规模数据集，专注于复杂城市环境中的多模态传感器融合任务。与 KITTI 数据集相比，nuScenes 数据集提供了更多的天气条件和更详细的标注信息，其包含 1 000 个场景，具体又分为 700 个训练场景、150 个验证场景和 150 个测试场景。nuScenes 数据集对 23 类目标进行了 3D 检测框标注，其有效检测距离为 100 m。

3.1.2 评价指标

KITTI 数据集主要针对汽车、骑行者和行人这 3 个类别进行检测效果评估，其评价指标主要有 $AP_{3D}$，$AP_{BEV}$ 以及平均方向相似度（average orientation similarity, AOS）。为了客观公正地比较不同模型的性能，我们与以往的工作[12,30]保持一致，采用 40 个召回位置计算 mAP，而不使用原先的 11 个召回位置。这种改进使得评估更加精确，并允许更细致的性能比较。此外，为了全面有效地进行扩展实验，我们还在 nuScenes 数据集上对比了本文方法和其他先进方法。nuScenes 数据集定义的主要评价指标为 mAP 和 nuScenes 检测分数（nuScenes detection score, NDS）。NDS 指标考虑了不同类型的检测错误，它是 mAP 和真阳性均值（mean true positive, mTP）的组合。

### 3.2 实验设置

3.2.1 实现细节

本文遵循以往的工作[45,58,69]，使用 MVP[44] 与 PENet[59] 方法从图像中生成虚拟点，而后将激光雷达点云和图像虚拟点云作为输入。动态缩放网络包含 4 个不同层级的动态卷积模块，前 2 个层级的特征维度分别是 16 和 32，后 2 个层级的特征维度均为 64。我们将 VirConv-T[45] 作为基线方法，同时使用与基线方法一致的损失函数。

3.2.2 训练详情

在 Pytorch 深度学习框架中，本文方法使用 2 块 NVIDIA GeForce RTX 3090 GPU 进行训练。我们对模型展开了 60 个轮次的训练，为了提高训练速度并降低训练成本，只对最后 15 个轮次进行数据集测验。实验学习率设置为 0.01，优化器为 Adam。为了扩展训练数据，增强动态缩放网络的鲁棒性，我们采用了一系列数据增强策略，包含旋转、平移以及翻转变换[23,45,69]。

### 3.3 实验结果

为了评估本文方法的性能，我们在 KITTI 和 nuScenes 这 2 个权威的主流数据集上进行了广泛实验，并与当前典型的 3D 目标检测方法进行了对比。实验结果表明，本文提出的 3D 动态缩放卷积和 2D 动态缩放卷积能够有效解决目标尺度差异问题。

3.3.1 在 KITTI 测试集上的结果

不同 3D 目标检测方法在 KITTI 测试集汽车类别上的性能对比如表 1 所示。从 $AP_{3D}$ 指标来看，在仅使用激光雷达的方法中，PV-RCNN[27] 的平均值最高，而多模态融合方法整体性能更优。尽管本文方法的 $AP_{3D}$ 指标在困难级别上的精度略逊于 VirConv-T[45]，为次优结果；但在简单和中等级别上的 $AP_{3D}$ 分别为 92.90% 和 86.52%，平均值为 86.87%，优于现有融合方法。从汽车 $AP_{BEV}$ 指标来看，本文方法在简单、中等和困难级别上的 $AP_{BEV}$ 分别为 96.63%，92.71%，89.74%，平均值为 93.03%，显著高于 VirConv-T，验证了所提方法在鸟瞰视角下对汽车检测的有效性。

不同 3D 目标检测方法在 KITTI 测试集行人、骑行者类别上的性能对比如表 2 所示。对于行人目标，尽管本文方法在中等级别上的 $AP_{3D}$ 低于 F-ConvNet[35]，为次优结果；但在简单和困难级别上的 $AP_{3D}$ 分别为 53.28% 和 41.81%，平均值为 46.68%，优于 CFMMFA[10] 等方法。对于骑行者目标，本文方法在简单、中等和困难级别上的 $AP_{3D}$ 分别为 81.53%，65.43%，58.21%，平均值为 68.39%，在多模态融合方法中表现最优，表明本文方法对行人、骑行者的检测鲁棒性较好。

3.3.2 在 KITTI 验证集上的结果

不同 3D 目标检测方法在 KITTI 验证集汽车类别上的性能对比如表 3 所示。从 $AP_{3D}$ 指标来看，尽管本文方法的 $AP_{3D}$ 在困难级别上略逊于 VirConv-T，为次优结果；但在简单和中等级别上的 $AP_{3D}$ 分别为 96.26% 和 90.64%，平均值为 91.66%，优于现有融合方法。从 $AP_{BEV}$ 指标来看，本文方法的平均值为 93.63%，高于 VirConv-T，进一步验证了本文方法的稳定性。

不同 3D 目标检测方法在 KITTI 验证集行人与骑行者类别上的性能对比如表 4 所示。对于行人目标，本文方法在 $AP_{3D}$ 上的平均值为 64.37%，高于 F-PointNet[34] 等方法，在多模态方法中排名第 1。对于骑行者目标，本文方法在 $AP_{3D}$ 上的平均值为 77.06%，在中等、困难级别上的 $AP_{3D}$ 分别为 72.04% 和 67.40%，高于 CasA-PV[23] 等单模态方法和 CFMMFA 等多模态方法，展现了复杂场景下对骑行者目标检测的有效性。

**Table 1 Comparison of 3D Object Detection Methods for Car Category on KITTI Test Set**

**表 1 3D 目标检测方法在 KITTI 测试集汽车类别上的对比**

| 方法 | 模态 | 汽车 $AP_{3D}$/% | | | | 汽车 $AP_{BEV}$/% | | | | 推理时间/ms |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 简单 | 中等 | 困难 | 平均值 | 简单 | 中等 | 困难 | 平均值 | |
| TANet[4] | L | 83.81 | 75.38 | 67.66 | 75.62 | | | | | |
| SVGA-Net[6] | L | 87.33 | 80.47 | 75.91 | 81.24 | | | | | |
| Point-GNN[29] | L | 88.33 | 79.47 | 72.29 | 80.03 | 93.11 | 89.17 | 83.90 | 88.73 | 600 |
| SA-SSD[28] | L | 88.75 | 79.79 | 74.16 | 80.90 | 95.03 | 91.03 | 85.96 | 90.67 | 40 |
| PV-RCNN[27] | L | 90.45 | 81.43 | 76.82 | 82.90 | 94.98 | 90.65 | 86.14 | 90.59 | 80 |
| F-PointNet[34] | L+C | 81.20 | 70.39 | 62.19 | 71.26 | 88.70 | 84.00 | 75.33 | 82.68 | 170 |
| PointPainting[41] | L+C | 82.11 | 71.70 | 67.08 | 73.63 | 92.45 | 88.11 | 83.36 | 87.97 | 400 |
| F-ConvNet[35] | L+C | 85.88 | 76.51 | 68.08 | 76.82 | 89.69 | 83.08 | 74.56 | 82.44 | 476 |
| CFMMFA[10] | L+C | 88.13 | 81.07 | 76.88 | 82.03 | 92.18 | 88.57 | 85.90 | 88.88 | 136 |
| DFIM[11] | L+C | 88.36 | 81.37 | 76.71 | 82.15 | 92.61 | 88.69 | 85.77 | 89.02 | 190 |
| SG-CAD[8] | L+C | 89.52 | 82.65 | 77.92 | 83.36 | 94.08 | 90.11 | 85.92 | 90.04 | 250 |
| EPNet[12] | L+C | 89.81 | 79.28 | 74.59 | 81.23 | 94.22 | 88.47 | 83.69 | 88.79 | 100 |
| VirConv-L[45] | L+C | 91.41 | 85.05 | 80.22 | 85.56 | 95.53 | 91.95 | 87.07 | 91.52 | 56 |
| VirConv-T[45] | L+C | 92.54 | 86.25 | **81.24** | 86.68 | 96.11 | 92.65 | 89.69 | 92.82 | 92 |
| 本文 | L+C | **92.90** | **86.52** | 81.19 | **86.87** | **96.63** | **92.71** | **89.74** | **93.03** | **39** |

注：黑体数值表示最优值，下划线数值表示次优值，“L”表示激光雷达，“L+C”表示激光雷达与相机融合。

**Table 2 Comparison of 3D Object Detection Methods for Pedestrian and Cyclist Categories on KITTI Test Set**

**表 2 3D 目标检测方法在 KITTI 测试集行人、骑行者类别上的对比**

| 方法 | 模态 | 行人 $AP_{3D}$/% | | | | 骑行者 $AP_{3D}$/% | | | | 推理时间/ms |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 简单 | 中等 | 困难 | 平均值 | 简单 | 中等 | 困难 | 平均值 | |
| SVGA-Net[6] | L | 48.48 | 40.39 | 37.92 | 42.26 | 78.58 | 62.28 | 54.88 | 65.25 | |
| Pointformer[24] | L | 50.67 | 42.23 | 39.60 | 44.23 | 75.01 | 59.80 | 53.99 | 62.93 | |
| Point-GNN[29] | L | 51.92 | 43.77 | 40.14 | 45.28 | 78.60 | 63.48 | 57.08 | 66.39 | 600 |
| PV-RCNN[27] | L | | | | | 78.60 | 63.71 | 57.65 | 66.65 | 80 |
| PointPainting[41] | L+C | 50.32 | 40.97 | 37.87 | 43.05 | 77.63 | 63.78 | 55.89 | 65.77 | 400 |
| AVOD-FPN[13] | L+C | 50.80 | 42.81 | 40.88 | 44.83 | 64.00 | 52.18 | 46.61 | 54.26 | 100 |
| F-PointNet[34] | L+C | 51.21 | 44.89 | 40.23 | 45.44 | 71.96 | 56.77 | 50.39 | 59.71 | 170 |
| CFMMFA[10] | L+C | 51.93 | 43.63 | 40.15 | 45.24 | 74.50 | 59.30 | 53.09 | 62.30 | 136 |
| F-ConvNet[35] | L+C | 52.37 | **45.61** | 41.49 | 46.49 | 79.58 | 64.68 | 57.03 | 67.10 | 476 |
| 本文 | L+C | **53.28** | 44.95 | **41.81** | **46.68** | **81.53** | **65.43** | **58.21** | **68.39** | **39** |

注：黑体数值表示最优值，下划线数值表示次优值，“L”表示激光雷达，“L+C”表示激光雷达与相机融合。

### 3.3.3 在 nuScenes 测试集上的结果

不同 3D 目标检测方法在 nuScenes 测试集上的性能对比如表 5 所示。从整体性能来看，基于激光雷达与相机融合的多模态方法在 NDS 和 mAP 指标上普遍优于仅激光雷达的方法，验证了多模态融合对于提升 3D 目标检测性能的有效性。在融合方法中，本文方法的综合性能较好，NDS 指标为 74.9%，mAP 指标为 71.8%，在所有对比方法中均为最高值。

具体来看，对于轿车、行人等高频目标，本文方法的 AP 分别为 88.5% 和 90.3%，均优于其他对比方法，体现了其在主流目标检测任务中的优势；对于施工车等低频目标，本文方法的 AP 为 38.1%，高于 CFMMFA 的 37.8%，在此类目标检测上达到了当前最高水平。对于卡车、公交车等中大型目标，本文方法的 AP 分别为 62.2% 和 72.0%，超过 CFMMFA 等方法且检测精度提升明显；对于摩托车、自行车等小目标，本文方法的 AP 分别为 74.9% 和 61.4%，显著高于 CFMMFA 等方法，表明其在小目标特征提取与识别

**Table 3 Comparison of 3D Object Detection Methods for Car Category on KITTI Validation Set**

**表 3 3D 目标检测方法在 KITTI 验证集汽车类别上的对比** %

| 方法 | 模态 | 汽车 $AP_{3D}$ | | | | 汽车 $AP_{BEV}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 简单 | 中等 | 困难 | 平均值 | 简单 | 中等 | 困难 | 平均值 |
| Point-GNN[29] | L | 87.89 | 78.34 | 77.38 | 81.20 | 89.82 | 88.31 | 87.16 | 88.43 |
| TANet[4] | L | 88.21 | 77.85 | 75.62 | 80.56 | | | | |
| SVGA-Net[6] | L | 90.59 | 80.23 | 79.15 | 83.32 | 90.27 | 89.16 | 88.11 | 89.18 |
| PV-RCNN[27] | L | 92.57 | 84.83 | 82.69 | 86.70 | 95.76 | 91.11 | 88.93 | 91.93 |
| CT3D[7] | L | 92.85 | 85.82 | 83.46 | 87.38 | 96.14 | 91.88 | 89.63 | 92.55 |
| F-PointNet[34] | L+C | 83.76 | 70.92 | 63.65 | 72.78 | 88.16 | 84.02 | 76.44 | 82.87 |
| PointPainting[41] | L+C | 88.38 | 77.74 | 76.76 | 80.96 | 90.19 | 87.64 | 86.71 | 88.18 |
| F-ConvNet[35] | L+C | 89.02 | 78.80 | 77.09 | 81.64 | 90.23 | 88.79 | 86.84 | 88.62 |
| DFIM[11] | L+C | 92.15 | 86.04 | 84.26 | 87.48 | 95.09 | 90.72 | 88.85 | 91.55 |
| CFMMFA[10] | L+C | 92.25 | 83.62 | 82.93 | 86.27 | 93.15 | 91.10 | 89.09 | 91.11 |
| EPNet[12] | L+C | 92.28 | 82.59 | 80.14 | 85.00 | 95.51 | 88.76 | 88.36 | 90.88 |
| SG-CAD[8] | L+C | 93.24 | 88.15 | 84.79 | 88.73 | 95.71 | 90.77 | 89.07 | 91.85 |
| VirConv-L[45] | L+C | 93.36 | 88.71 | 85.83 | 89.30 | 96.00* | 91.62* | 91.20* | 92.94* |
| VirConv-T[45] | L+C | 95.81 | 90.29 | **88.10** | 91.40 | 96.26* | 91.71* | 91.31* | 93.09* |
| 本文 | L+C | **96.26** | **90.64** | 88.09 | **91.66** | **96.88** | **92.24** | **91.77** | **93.63** |

注：黑体数字表示最优值，下划线数值表示次优值，“L”表示激光雷达，“L+C”表示激光雷达与相机融合，“*”表示自测值。

**Table 4 Comparison of 3D Object Detection Methods for Pedestrian and Cyclist Categories on KITTI Validation Set**

**表 4 3D 目标检测方法在 KITTI 验证集行人、骑行者类别上的对比** %

| 方法 | 模态 | 行人 $AP_{3D}$ | | | | 骑行者 $AP_{3D}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 简单 | 中等 | 困难 | 平均值 | 简单 | 中等 | 困难 | 平均值 |
| VoxelNet[14] | L | 57.86 | 53.42 | 48.87 | 53.38 | 67.17 | 47.65 | 45.11 | 53.31 |
| CT3D[7] | L | 65.73 | 58.56 | 53.04 | 59.11 | **91.99** | 71.60 | 67.34 | 76.98 |
| CasA-PV[23] | L | 68.90 | 59.86 | 53.66 | 60.81 | 90.95 | 70.69 | 65.82 | 75.82 |
| TANet[4] | L | 70.80 | 63.45 | **58.22** | 64.16 | 85.98 | 64.95 | 60.40 | 70.44 |
| AutoAlign[36] | L+C | 54.87 | 48.53 | 44.61 | 49.34 | 82.25 | 63.40 | 58.89 | 68.18 |
| CFMMFA[10] | L+C | 66.94 | 60.82 | 56.09 | 61.28 | 88.62 | 70.01 | 65.95 | 74.86 |
| PointPainting[41] | L+C | 69.38 | 61.67 | 54.58 | 61.88 | 85.21 | 71.62 | 66.98 | 74.60 |
| SG-CAD[8] | L+C | 69.54 | **64.81** | 55.82 | 63.39 | 83.52 | 70.31 | 65.79 | 73.21 |
| F-PointNet[34] | L+C | 70.00 | 61.32 | 53.59 | 61.64 | 77.15 | 56.49 | 53.37 | 62.34 |
| 本文 | L+C | **71.38** | 64.12 | 57.60 | **64.37** | 91.73 | **72.04** | **67.40** | **77.06** |

注：黑体数值表示最优值，下划线数值表示次优值，“L”表示激光雷达，“L+C”表示激光雷达与相机融合。

上的有效性。此外，本文方法在交通锥和障碍物的检测上也表现突出，AP 分别为 87.2% 和 80.4%，均为所有方法中的最优值，验证了复杂场景下所提方法对多类别目标检测的鲁棒性。

3.3.4 在效率指标上的结果

为评估本文方法的有效性，尤其是面向自动驾驶场景的实时性需求，本节从计算复杂度与推理速度这 2 个方面对模型效率进行定量分析。具体地，分析了本文方法与典型方法在参数量、推理时间与推理速度方面的性能，对比结果如表 6 所示。

根据表 6 可知，本文方法在保持高检测精度的同时展现出显著的计算效率优势。在参数量方面，本文方法参数量仅为 $12.9\times10^6$，低于绝大多数对比方法，甚至低于轻量化的 PV-RCNN[27] 与 Focals Conv[51] 方法。这表明动态缩放卷积及其双分支结构在实现自适应特征提取的同时并未引入冗余参数。在推理时间和推理速度方面，本文方法的推理时间为 39 ms，推理速度为 25.6 FPS，在 10 种对比方法中表现突出，

**Table 5 Comparison of 3D Object Detection Methods on nuScenes Test Set**

**表 5 3D 目标检测方法在 nuScenes 测试集上的对比** %

| 方法 | 模态 | NDS | mAP | AP | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 轿车 | 卡车 | 拖车 | 行人 | 公交车 | 施工车 | 摩托车 | 自行车 | 交通锥 | 障碍物 |
| CenterPoint[5] | L | 65.5 | 58.0 | 84.6 | 51.0 | 53.2 | 83.4 | 60.2 | 17.5 | 53.7 | 28.7 | 76.7 | 70.9 |
| Focals Conv[51] | L | 70.0 | 63.8 | 86.7 | 56.3 | 59.5 | 87.5 | 67.7 | 23.8 | 64.5 | 36.3 | 81.4 | 74.1 |
| VoxelNext[30] | L | 70.0 | 64.5 | 84.6 | 53.0 | 55.8 | 85.8 | 64.7 | 28.7 | 73.2 | 45.7 | 79.0 | 74.6 |
| TransFusion-L[9] | L | 70.2 | 65.5 | 86.2 | 56.7 | 58.8 | 86.1 | 66.3 | 28.2 | 68.3 | 44.2 | 82.0 | 78.2 |
| PointPainting[41] | L+C | 58.1 | 46.4 | 77.9 | 35.8 | 37.3 | 73.3 | 36.1 | 15.8 | 41.5 | 24.1 | 62.4 | 60.2 |
| MVP[44] | L+C | 70.5 | 66.4 | 86.8 | 58.5 | 57.3 | 89.1 | 67.4 | 26.1 | 70.0 | 49.3 | 85.0 | 74.8 |
| LargeKernel3D[52] | L+C | 70.6 | 65.4 | 85.5 | 53.8 | 59.5 | 85.9 | 64.4 | 29.7 | 72.7 | 46.8 | 79.9 | 75.5 |
| TransFusion[9] | L+C | 71.7 | 68.9 | 87.1 | 60.0 | 60.8 | 88.4 | 68.3 | 33.1 | 73.6 | 52.9 | 86.7 | 78.1 |
| CFMMFA[10] | L+C | 73.2 | 70.7 | 88.1 | 60.1 | **64.1** | 90.0 | 70.5 | 37.8 | 74.6 | 54.9 | 86.8 | 80.2 |
| 本文 | L+C | **74.9** | **71.8** | **88.5** | **62.2** | 62.6 | **90.3** | **72.0** | **38.1** | **74.9** | **61.4** | **87.2** | **80.4** |

注：黑体数值表示最优值，下划线数值表示次优值，“L”表示激光雷达，“L+C”表示激光雷达与相机融合。

**Table 6 Comparison of 3D Object Detection Methods on Efficiency Metrics**

**表 6 3D 目标检测方法在效率指标上的对比**

| 方法 | 参数量 | 推理时间/ms | 推理速度/FPS |
|---|---|---|---|
| Focals Conv[51] | $13.4\times10^6$ | 100 | 10.0 |
| DeepInteraction[70] | $57.8\times10^6$ | 204 | 4.9 |
| RoboFusion-B[71] | $81.0\times10^6$ | 286 | 3.5 |
| BEVFormer[72] | $68.7\times10^6$ | 588 | 1.7 |
| SparseBEV[73] | $70.0\times10^6$ | 43 | 23.5 |
| LargeKernel3D[52] | $40.2\times10^6$ | 145 | 6.9 |
| SG-CAD[8] | $32.0\times10^6$ | 250 | 4.0 |
| PV-RCNN[27] | $13.1\times10^6$ | 80 | 12.5 |
| TransFusion[9] | $36.9\times10^6$ | 161 | 6.2 |
| 本文 | **$12.9\times10^6$** | **39** | **25.6** |

注：黑体数值表示最优值，下划线数值表示次优值。

明显优于 SG-CAD[8], TransFusion[9], LargeKernel3D[52] 等多模态融合方法。这主要得益于动态缩放卷积的自适应采样与矩阵运算优化操作减少了点云计算开销。同时，双分支之间的权重共享进一步降低了计算负担。

### 3.4 可视化结果

本文还在 KITTI 验证集上进行了目标检测结果的可视化。如图 5 所示，第 1 行和第 3 行呈现的是融合点云，融合点云由激光雷达点云与图像虚拟点云融合得到；而第 2 行和第 4 行呈现的是相机图像。对于每一列内容，第 1 行融合点云和第 2 行相机图像分别从 3D 和 2D 维度反映同一场景，第 3 行融合点云和第 4 行相机图像分别从 3D 和 2D 维度反映同一场景。

观察来看，无论是简单场景还是复杂场景，本文模型的预测结果均与真实结果的重合度较高，能够准确地检测出目标。从左到右观察不同列的场景，随着场景从开阔的道路交叉口、相对空旷的街道逐渐过渡到车辆密集的街道、高速公路，目标分布及环境逐渐复杂，但模型仍能有效检测场景中的目标，展现了较好的鲁棒性与适应性。

### 3.5 消融实验

#### 3.5.1 卷积性能分析

在 KITTI 验证集上，我们对 3D 动态缩放卷积和 2D 动态缩放卷积的有效性进行了消融实验。

在 KITTI 验证集的汽车类别上，由表 7 可知，对于不同难度级别下的 $AP_{3D}$ 与 $AP_{BEV}$ 指标，基线方法在加入 2D 动态缩放卷积后性能略有提升；加入 3D 动态缩放卷积后性能提升更明显；二者共同启用时性能最优。其中，在简单、中等和困难级别上的 $AP_{3D}$ 分别为 96.26%, 90.64%, 88.09%，在简单、中等和困难级别上的 $AP_{BEV}$ 分别为 96.88%, 92.24%, 91.77%。这表明在汽车检测上，3D 动态缩放卷积和 2D 动态缩放卷积对模型性能均有正向作用，且 3D 动态缩放卷积的提升效果优于 2D 动态缩放卷积，二者结合时能产生协同效应，在提升模型检测性能方面具备有效性和互补性。

在 KITTI 验证集的行人类别上，由图 6(a)可知，对于不同难度级别下的 $AP_{3D}$ 与 $AP_{BEV}$ 指标，基线方法最低；加入 2D 动态缩放卷积后性能有一定提升；加入 3D 动态缩放卷积后性能提升幅度更大；二者共同启用时性能最优。这表明 3D 动态缩放卷积和 2D 动态缩放卷积对行人检测均有正向作用，且 3D 动态

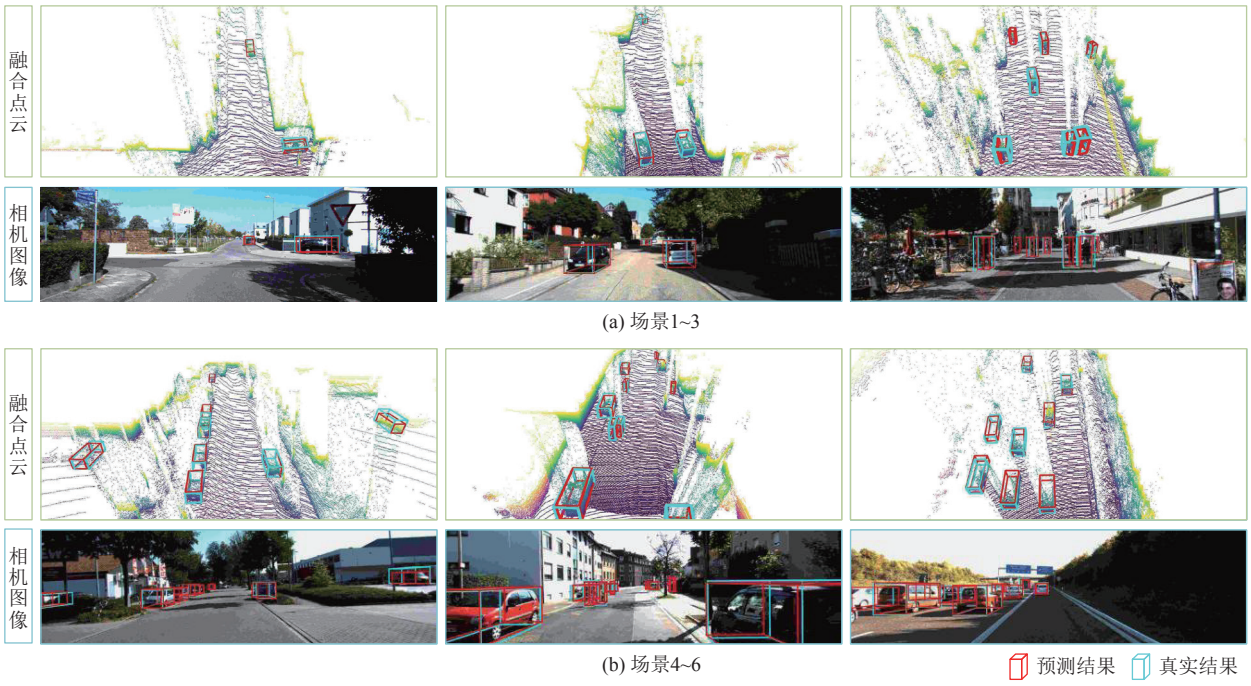



Fig. 5 Object detection results under different scenarios

图 5 不同场景下的目标检测结果

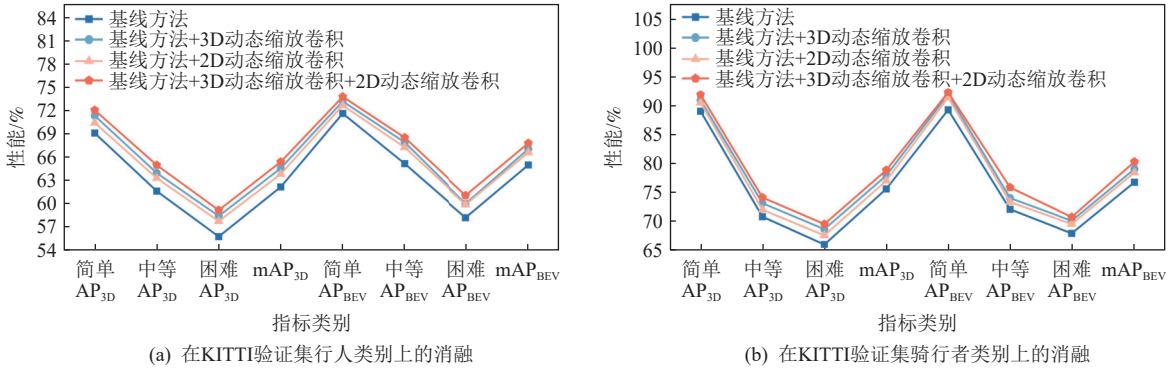
**Table 7 Ablation Study for Car Category on KITTI Validation Set**

**表 7 在 KITTI 验证集汽车类别上的消融实验** %

| 基线方法 | 3D 动态缩放卷积 | 2D 动态缩放卷积 | 汽车 $AP_{3D}$ | | | | 汽车 $AP_{BEV}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 简单 | 中等 | 困难 | 平均值 | 简单 | 中等 | 困难 | 平均值 |
| √ | × | × | 95.81 | 90.29 | 88.10 | 91.40 | 96.26 | 91.71 | 91.31 | 93.09 |
| √ | √ | × | 96.20 | 90.55 | 88.06 | 91.60 | 96.74 | 92.13 | 91.69 | 93.52 |
| √ | × | √ | 96.00 | 90.45 | 88.12 | 91.52 | 96.51 | 92.18 | 91.52 | 93.40 |
| √ | √ | √ | 96.26 | 90.64 | 88.09 | 91.66 | 96.88 | 92.24 | 91.77 | 93.63 |

缩放卷积提升更显著。在 KITTI 验证集的骑行者类别上，由图 6(b)可知，对于不同难度级别下的 $AP_{3D}$ 与 $AP_{BEV}$ 指标，基线方法最低；加入 2D 动态缩放卷积后性能略有改善；加入 3D 动态缩放卷积后性能提升更为明显；二者同时启用时性能最佳。这表明 3D 动态缩放卷积和 2D 动态缩放卷积的引入对汽车、行人、骑行者等不同尺度目标的检测性能均有显著提升。

值得注意的是，在尺度更小、形态变化更明显的行人、骑行者类别上，性能提升尤为突出。这验证了动态缩放卷积机制的核心优势：通过自适应调整感受野内有效采样位置与采样点密度，能够精准匹配不同目标的固有尺度与空间结构特征。对于汽车等大尺度目标，动态缩放卷积能够有效扩大感受野以整合其宏观轮廓信息；对于行人、骑行者等小尺度目标，则能收缩并聚焦于局部高分辨特征，从而缓解因固定卷积核导致的特征模糊或空间细节丢失问题。因此，本文方法并非仅对特定尺度目标有效，而是通过动态缩放机制系统增强模型对不同尺度目标的特




(a) 在KITTI验证集行人类别上的消融 (b) 在KITTI验证集骑行者类别上的消融

Fig. 6 Comparative line plots of the ablation study

图 6 消融实验对比折线图

征表示能力, 为应对现实场景中普遍存在的尺度差异问题提供了有效解决方案。

综合来看, 引入 3D 动态缩放卷积和 2D 动态缩放卷积能够不同程度地提升模型的检测性能, 且二者结合时检测效果达到最佳, 这表明 2 种动态缩放卷积在特征提取和目标检测过程中具有互补作用, 能够提高模型对不同难度目标的检测精度。

### 3.5.2 关键参数分析

为深入探究动态缩放卷积的内部机制, 本节从卷积层数配置和尺度学习范围这 2 个方面, 对关键参数进行了消融实验。

在动态缩放卷积层数方面, 利用 KITTI 验证集分析了不同层数配置对检测性能的影响, 如表 8 所示。对于 3D 动态缩放卷积, 当层数从 1 层增加到 2 层时, 汽车类别的 $AP_{3D}$ 平均值从 91.45% 提升至 91.60%, 行人类别的 $AP_{3D}$ 平均值从 64.04% 提升至 64.18%, 骑行者类别的 $AP_{3D}$ 平均值从 76.63% 提升至 76.77%, 这表明适度增加层数有助于捕获更丰富的目标特征。然而, 当层数增加至 3 层及以上时, 3D 动态缩放卷积的性能逐渐下降, 说明过深的网络结构会导致过拟合和梯度传播困难。而 2D 动态缩放卷积也呈现出相似的规律, 即配置 2 层卷积时对不同尺度目标的检测效果最优。因此, 本文采用 2 层 3D 动态缩放卷积和 2 层 2D 动态缩放卷积的配置方案。

在卷积核尺度学习范围方面, 利用 KITTI 验证集分析了不同卷积核尺度范围对性能的影响, 如图 7 所示。当 3D 卷积核的高度、宽度和深度学习范围逐步从 1~3 扩展至 1~15 时, 其性能有所提升并在 1~15 时达到最大; 但从 1~15 扩展至 1~27 时, 3D 卷积核的 $mAP_{3D}$ 超出最优设置并开始下降。这种现象出现的原因是, 较小的尺度学习范围会导致采样点分布密集, 从而捕获过多噪声并造成冗余; 而较大的尺度学习范围则会因采样点分布稀疏而影响性能。2D 动态缩放卷积也呈现出相似的规律, 即卷积核尺度学习范围为 1~15 时达到性能与计算效率间的最佳权衡。

**Table 8 Ablation Study on Dynamic Scaling Convolution Layers**

**表 8 在动态缩放卷积层上的消融实验** %

| 类别 | 层数 | 汽车 $AP_{3D}$ | | | | 行人 $AP_{3D}$ | | | | 骑行者 $AP_{3D}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 简单 | 中等 | 困难 | 平均值 | 简单 | 中等 | 困难 | 平均值 | 简单 | 中等 | 困难 | 平均值 |
| 3D 动态缩放卷积 | 1 | 95.97 | 90.42 | 87.97 | 91.45 | 71.18 | 63.86 | 57.07 | 64.04 | 90.78 | 71.97 | 67.13 | 76.63 |
| | 2 | **96.20** | **90.55** | **88.06** | **91.60** | **71.25** | **63.91** | **57.39** | **64.18** | **90.89** | **72.15** | **67.28** | **76.77** |
| | 3 | 95.89 | 90.24 | 87.78 | 91.30 | 71.20 | 63.81 | 56.95 | 63.99 | 90.84 | 72.01 | 67.21 | 76.69 |
| | 4 | 95.81 | 90.2 | 87.69 | 91.23 | 71.09 | 63.72 | 56.87 | 63.89 | 90.63 | 71.76 | 66.98 | 76.46 |
| | 5 | 95.73 | 90.11 | 87.43 | 91.09 | 70.97 | 63.65 | 56.81 | 63.81 | 90.55 | 71.62 | 66.74 | 76.30 |
| 2D 动态缩放卷积 | 1 | 95.88 | 90.35 | 87.92 | 91.38 | 70.23 | 63.18 | 56.69 | 63.37 | 90.41 | 70.98 | 66.52 | 75.97 |
| | 2 | **96.00** | **90.45** | **88.12** | **91.52** | **70.34** | **63.25** | **56.73** | **63.44** | **90.45** | **71.03** | **66.61** | **76.03** |
| | 3 | 95.79 | 90.22 | 87.77 | 91.26 | 70.31 | 63.24 | 56.64 | 63.40 | 90.38 | 70.87 | 66.36 | 75.87 |
| | 4 | 95.76 | 90.14 | 87.61 | 91.17 | 70.12 | 63.03 | 56.41 | 63.19 | 90.32 | 70.68 | 66.15 | 75.72 |
| | 5 | 95.68 | 90.03 | 87.34 | 91.01 | 70.01 | 62.81 | 56.13 | 62.98 | 90.11 | 70.42 | 65.93 | 75.49 |

注: 黑体数值表示最优值。

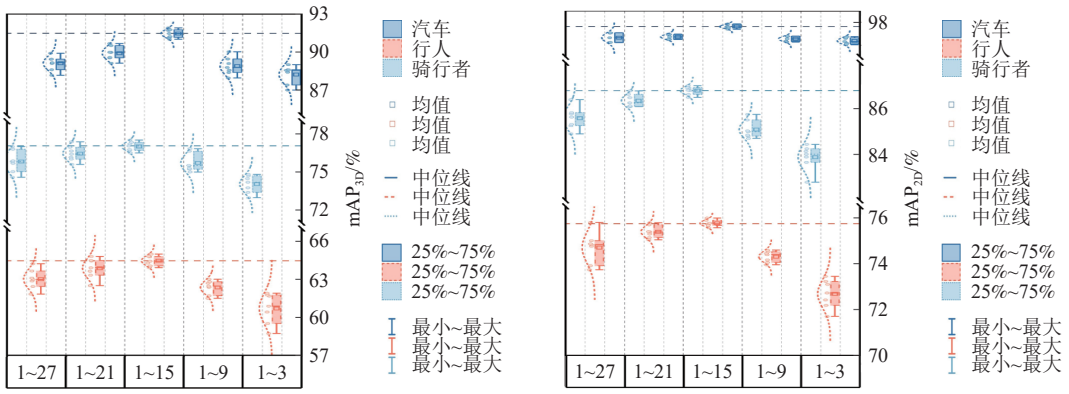
### 3.5.3 检测效果分析

为验证所提方法解决目标尺度差异问题的有效性, 我们基于 KITTI 验证集评估了不同距离下的 $AP_{3D}$。具体地, 在 0~15 m, 15~30 m, 30~45 m, 45 m 以上这 4 个递进的距离区间上, 对汽车、行人及骑行者目标进行了检测, 结果如图 8 所示。在远距离场景中, 汽车、行人及骑行者目标检测均实现了更出色的性能提升, 这是因为动态缩放卷积能够应对不同检测距离带来的尺度差异挑战, 提取远距离稀疏目标的 3D 特征。

## 3.6 结果分析与讨论

尽管本文方法在 KITTI 与 nuScenes 数据集上的综合性能优于对比方法, 但在个别具体指标上仍有提升空间, 如 KITTI 测试集汽车类别的困难级别 $AP_{3D}$ 和行人类别的中等级别 $AP_{3D}$ 等。我们分析其原因主要包含 2 个方面。

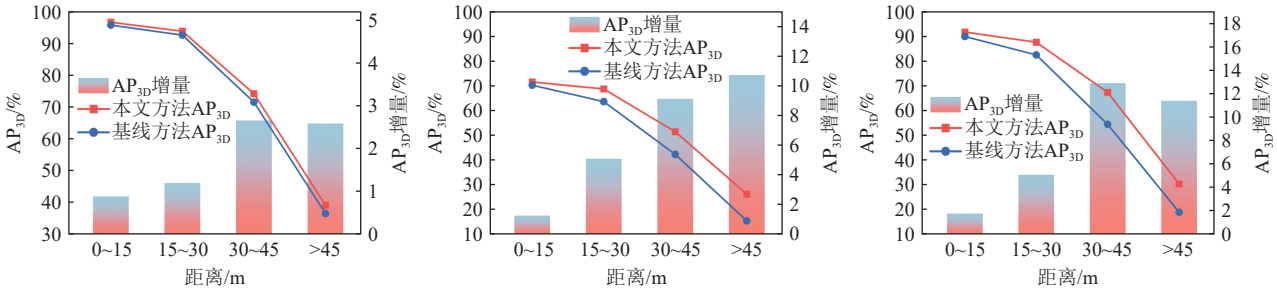
1)极端尺度与稀疏性挑战。对于远距离或尺度极小的目标, 在点云极度稀疏区域, 动态缩放卷积的自适应采样机制可能存在细微的优化空间。研究表明, 点云空间分布不均匀会显著影响特征提取质量,




(a) 在3D动态缩放卷积核上的消融　(b) 在2D动态缩放卷积核上的消融

Fig. 7　Grouped box plots of the ablation study

图 7　消融实验分组箱线图




(a) 不同距离下的汽车类别检测　(b) 不同距离下的行人类别检测　(c) 不同距离下的骑行者类别检测

Fig. 8　$AP_{3D}$ and performance improvement at different distances

图 8　不同距离下的 $AP_{3D}$ 与性能提升

特别是在远距离目标检测中，稀疏点云难以提供足够的几何信息[58]。因此，可以考虑多分辨率空间金字塔结合稀疏卷积等方法[55,74]，改善远距离目标检测的表现。

2)模态互补的提升瓶颈。在强光、运动模糊等图像质量退化的场景下，图像虚拟点云的生成质量受限，影响了多模态融合的最终效益。这是因为图像质量退化会直接降低深度估计精度[75]，进而影响虚拟点云的空间准确性。尤其在遭受极端光照或攻击时，基于图像的特征提取会出现显著的性能下降，导致跨模态对齐误差增大[76-78]。此外，相机运动模糊会破坏时空一致性，使得多模态融合难以充分发挥互补优势[79]。

针对这些挑战，未来可以进一步优化自适应尺度感知与跨模态交互机制，提升 3D 目标检测的鲁棒性。

## 4　结　　论

本文针对多模态 3D 目标检测中因目标尺度差异显著以及标准卷积感受野局限导致的性能瓶颈，设计了适用于 3D 空间和 2D 空间的动态缩放卷积机制，能够动态调整卷积核的采样位置与采样点数量，实现多尺度特征提取与跨模态语义对齐。基于 3D 动态缩放卷积和 2D 动态缩放卷积，我们构建了动态缩放网络并与体素网络共同构成双分支网络架构。该架构能够协同处理激光雷达点云与图像虚拟点云，在保持特征一致性的同时增强复杂场景下的感知鲁棒性。在 KITTI 与 nuScenes 数据集上的实验表明，所提方法在多个目标类别上均达到先进性能，尤其在小目标与复杂场景下表现出优良的鲁棒性。下一步我们将继续探索生物视觉系统中的自适应感知机制，为构建更具解释性和通用性的多模态感知模型提供新思路。

**作者贡献声明**：纪文宇主导研究设计、核心实验及论文撰写；苗壮负责关键算法优化验证与实验方案细化；张启阳负责实验操作、数据收集与分析；崔小璐参与论文修改并提出改进意见；李唯暄协助数据预处理、辅助实验及结果讨论；李阳统筹研究进度与团队协作，确定选题并进行指导。

## 参 考 文 献


[1] Jing Hao, Wang Anhong, Zhao Lijun, et al. Boosting 3D object detection with semantic-aware multi-branch framework[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(6): 5697−5710

[2] Qi Yafei, Yang Menghao, Zhang Yongmin, et al. SelfDN: Adaptive self-denoising for multi-view 3D object detection[J]. Knowledge-Based Systems, 2025, 324: 113816

[3] Fan Lue, Wang Feng, Wang Naiyan, et al. FSDv2: Improving fully sparse 3D object detection with virtual voxels[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 47(2): 1279−1292

[4] Liu Zhe, Zhao Xin, Huang Tengteng, et al. TANet: Robust 3D object detection from point clouds with triple attention[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11677−11684

[5] Yin Tianwei, Zhou Xingyi, Krahenbuhl P. Center-based 3D object detection and tracking[C]//Proc of the 38th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 11784−11793

[6] He Qingdong, Wang Zhenning, Zeng Hao, et al. SVGA-Net: Sparse voxel-graph attention network for 3D object detection from point clouds[C]//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 870−878

[7] Sheng Hualian, Cai Sijia, Liu Yuan, et al. Improving 3D object detection with channel-wise Transformer[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 2743−2752

[8] Mushtaq H, Deng Xiaoheng, Alizadehsani R, et al. SC3D: Semantic-guided and class-adaptive cross-domain fusion for 3D object detection in autonomous vehicles[J]. Expert Systems with Applications, 2025, 268: 126359

[9] Bai Xuyang, Hu Zeyu, Zhu Xinge, et al. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers[C]//Proc of the 39th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 1090−1099

[10] Yu Fengning, Lian Jing, Li Linhui, et al. Cascade fusion of multi-modal and multi-source feature fusion by the attention for three-dimensional object detection[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108124

[11] Chen Mu, Liu Pengfei, Zhao Huaici. LiDAR-camera fusion: Dual transformer enhancement for 3D object detection[J]. Engineering Applications of Artificial Intelligence, 2023, 120: 105815

[12] Huang Tengteng, Liu Zhe, Chen Xiwu, et al. EPNet: Enhancing point features with image semantics for 3D object detection[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 35−52

[13] Ku J, Mozifian M, Lee J, et al. Joint 3D proposal generation and object detection from view aggregation[C]//Proc of the 2018 IEEE/RSJ Int Conf on Intelligent Robots and Systems. Piscataway, NJ: IEEE: 1−8

[14] Zhou Yin, Tuzel O. VoxelNet: End-to-end learning for point cloud based 3D object detection[C]//Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4490−4499

[15] Qi Yaolei, He Yuting, Qi Xiaoming, et al. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 6070−6079

[16] Zhao Tianchen, Ning Xuefei, Hong Ke, et al. Ada3D: Exploiting the spatial redundancy with adaptive inference for efficient 3D object detection[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 17728−17738

[17] Li Yuxuan, Li Xiang, Dai Yimian, et al. LSKNet: A foundation lightweight backbone for remote sensing[J]. International Journal of Computer Vision, 2025, 133(3): 1410−1431

[18] Lu Jiahao, Deng Jiacheng. Relation3D: Enhancing relation modeling for point cloud instance segmentation[C]//Proc of the 42nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2025: 8889−8899

[19] Finder S E, Amoyal R, Treister E, et al. Wavelet convolutions for large receptive fields[C]//Proc of the 18th European Conf on Computer Vision. Berlin: Springer, 2024: 363−380

[20] Sun Z, Chen Y J, Yang Y H, et al. Machine learning modelling for multi-order human visual motion processing[J]. Nature Machine Intelligence, 2025, 7(7): 1037−1052

[21] Harrison W J, Bays P M, Rideaux R. Neural tuning instantiates prior expectations in the human visual system[J]. Nature Communications, 2023, 14: 5320

[22] Zapp S J, Nitsche S, Gollisch T. Retinal receptive-field substructure: Scaffolding for coding and computation[J]. Trends in Neurosciences, 2022, 45(6): 430−445

[23] Wu Hai, Deng Jinhao, Wen Chenglu, et al. CasA: A cascade attention network for 3D object detection from LiDAR point clouds[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1−11

[24] Pan Xuran, Xia Zhuofan, Song Shiji, et al. 3D object detection with pointformer[C]//Proc of the 38th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 7463−7472

[25] Han Bing, Zhang Xinyun, Ren Shuang. A survey of convolutional operations based on 3D point clouds[J]. Journal of Computer Research and Development, 2023, 60(4): 873−902 (in Chinese)
(韩冰, 张鑫云, 任爽. 基于三维点云的卷积运算综述[J]. 计算机研究与发展, 2023, 60(4): 873−902)

[26] Lu Sidi, He Yuankai, Shi Weisong. Vehicle computing: A new computing paradigm in the era of autonomous driving[J]. Journal of Computer Research and Development, 2025, 62(1): 2−21 (in Chinese)
(鲁思迪, 何元恺, 施巍松. 车计算: 自动驾驶时代的新型计算范式[J]. 计算机研究与发展, 2025, 62(1): 2−21)

[27] Shi Shaoshuai, Guo Chaoxu, Jiang Li, et al. PV-RCNN: Point-voxel feature set abstraction for 3D object detection[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10529−10538

[28] He Chenhang, Zeng Hui, Huang Jianqiang, et al. Structure aware single-stage 3D object detection from point cloud[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 11873−11882

[29] Shi Weijing, Rajkumar R. Point-GNN: Graph neural network for 3D object detection in a point cloud[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 1711−1719

[30] Chen Yukang, Liu Jianhui, Zhang Xiangyu, et al. VoxelNext: Fully sparse voxelnet for 3D object detection and tracking[C]//Proc of the 40th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 21674−21683

[31] Li Jiaojiao, Sun Hongyan, Dong Yu, et al. A survey of 3D point cloud processing based on deep learning[J]. Journal of Computer Research and Development, 2022, 59(5): 1160−1179 (in Chinese)
(李娇娇, 孙红岩, 董雨, 等. 基于深度学习的 3 维点云处理综述[J]. 计算机研究与发展, 2022, 59(5): 1160−1179)

[32] Ji Wenyu, Li Yang, Wang Jiabao, et al. A review of 3D object detection based on LiDAR-camera fusion[J]. Computer Science, 2026, 53(6): 214−231 (in Chinese)
(纪文宇, 李阳, 王家宝, 等. 基于激光雷达与相机融合的 3D 目标检测综述[J]. 计算机科学, 2026, 53(6): 214−231)

[33] Ji Wenyu, Li Yang, Zhang Qiyang, et al. Gradient manifold density fusion for 3D object detection[C]//Proc of the 21st Int Conf on Intelligent Computing. Berlin: Springer, 2025: 195−206

[34] Qi C R, Liu Wei, Wu Chenxia, et al. Frustum pointnets for 3D object detection from RGB-D data[C]//Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 918−927

[35] Wang Zhixin, Jia Kui. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection[C]//Proc of the 2019 IEEE/RSJ Int Conf on Intelligent Robots and Systems. Piscataway, NJ: IEEE, 2019: 1742−1749

[36] Chen Zehui, Li Zhenyu, Zhang Shiquan, et al. AutoAlign: Pixel-instance feature aggregation for multi-modal 3D object detection[J]. arXiv preprint, arXiv: 2201.06493, 2025

[37] Wang Lin, Sun Shiliang, Jing Zhao. VirpNet: A multi-modal virtual point generation network for 3D object detection[J]. IEEE Transactions on Multimedia, 2024, 26: 10597−10609

[38] Yin Junbo, Shen Jianbing, Chen Runnan, et al. Is-fusion: Instance-scene collaborative fusion for multi-modal 3D object detection[C]// Proc of the 41st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 14905−14915

[39] Pang Su, Morris D, Radha H. Fast-CLOCs: Fast camera-LiDAR object candidates fusion for 3D object detection[C]//Proc of the 22nd IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2022: 187−196

[40] Zhang Tingyu, Liang Zhigang, Yang Yanzhao, et al. Contrastive late fusion for 3D object detection[J]. IEEE Transactions on Intelligent Vehicles, 2025, 10(5): 3442−3457

[41] Vora S, Lang A H, Helou B, et al. PointPainting: Sequential fusion for 3D object detection[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 4604−4612

[42] Wang Haiyang, Hao Tang, Shi Shaoshuai, et al. UniTR: A unified and efficient multi-modal transformer for bird's-eye-view representation[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 6792−6802

[43] Xie Yichen, Xu Chenfeng, Rakotosaona M J, et al. SparseFusion: Fusing multi-modal sparse representations for multi-sensor 3D object detection[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 17591−17602

[44] Yin Tianwei, Zhou Xingyi, Krähenbühl P. Multi-modal virtual point 3D detection[C]//Proc of the 35th Int Conf on Neural Information Processing Systems. https://dl.acm.org/doi/10.5555/3540261.3541522

[45] Wu Hai, Wen Chenglu, Shi Shaoshuai, et al. Virtual sparse convolution for multi-modal 3D object detection[C]//Proc of the 40th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 21653−21662

[46] Wu Hai, Zhao Shijia, Huang Xun, et al. Unsupervised 3D object detection by commonsense clue[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(12): 11720−11737

[47] Guo Wenxuan, Xu Xiuwei, Wang Ziwei, et al. Text-guided sparse voxel pruning for efficient 3D visual grounding[C]//Proc of the 42nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2025: 3666−3675

[48] Chen Zixuan, He Zewei, Lu Z M. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention[J]. IEEE Transactions on Image Processing, 2024, 33: 1002−1015

[49] Liu Hui, Jia Chen, Shi Fan, et al. SCSegamba: Lightweight structure-aware vision mamba for crack segmentation in structures[C]//Proc of the 42nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2025: 29406−29416

[50] Liu Xinyue, Yin Hailian, Zang Yalei, et al. Sequential recommendation based on graph interpolation and deformable convolutional networks[J]. Journal of Computer Research and Development, 2025, 62(10): 2583−2594 (in Chinese)
(刘昕悦, 尹海莲, 臧亚磊, 等. 基于图插值和可变形卷积网络的序列推荐[J]. 计算机研究与发展, 2025, 62(10): 2583−2594)

[51] Chen Yukang, Li Yanwei, Zhang Xiangyu, et al. Focal sparse convolutional networks for 3D object detection[C]//Proc of the 39th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 5428−5437

[52] Chen Yukang, Liu Jianhui, Zhang Xiangyu, et al. LargeKernel3D: Scaling up kernels in 3D sparse CNNs[C]//Proc of the 40th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 13488−13498

[53] Liu Zhijian, Tang Haotian, Zhao Shengyu, et al. PVNAS: 3D neural architecture search with point-voxel convolution[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 8552−8568

[54] Wei Tao, Tian Yonghong, Wang Yaowei, et al. Optimized separable convolution: Yet another efficient convolution operator[J]. AI Open, 2022, 3: 162−171

[55] Yang Jiannan, Liu Shuangli, Wu Jingjun, et al. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target

detection[C]//Proc of the 39th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025: 9202−9210

[56] Cai Zhicheng, Ding Xiaohan, Shen Qiu, et al. RefConv: Reparameterized refocusing convolution for powerful convnets[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(6): 11617−11631

[57] Wu Hai, Wen Chenglu, Li Wei, et al. Transformation-equivariant 3D object detection for autonomous driving[C]//Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 2795−2802

[58] Yu Zhuochen, Qiu Bijie, Khong A W H. ViKIENet: Towards efficient 3D object detection with virtual key instance enhanced network[C]//Proc of the 37th Computer Vision and Pattern Recognition Conf. Piscataway, NJ: IEEE, 2025: 11844−11853

[59] Hu Mu, Wang Shuling, Li Bin, et al. PENet: Towards precise and efficient image guided depth completion[C]//Proc of the 2021 IEEE Int Conf on Robotics and Automation. Piscataway, NJ: IEEE, 2021: 13656−13662

[60] Yan Yan, Mao Yuxing, Li Bo. Second: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337

[61] Graham B, Engelcke M, Van D M L. 3D semantic segmentation with submanifold sparse convolutional networks[C]//Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 9224−9232

[62] Li Chengye, Zhang Zhen, Liang Zheheng, et al. A survey of object detection models[J]. Journal of Computer Research and Development, 2025, 62(12): 2895−2928 (in Chinese)
(李承烨, 张震, 梁哲恒, 等. 目标检测模型综述[J]. 计算机研究与发展, 2025, 62(12): 2895−2928)

[63] Wang Xueyang, Zheng Zhixin, Shao Jiandong, et al. Adaptive rectangular convolution for remote sensing pansharpening[C]//Proc of the 42nd Computer Vision and Pattern Recognition Conf. Piscataway, NJ: IEEE, 2025: 17872−17881

[64] Tchapmi L, Choy C, Armeni I, et al. Segcloud: Semantic segmentation of 3D point clouds[C]//Proc of the 5th Int Conf on 3D Vision. Piscataway, NJ: IEEE, 2017: 537−547

[65] Meijering E. A chronology of interpolation: From ancient astronomy to modern signal and image processing[J]. Proceedings of the IEEE, 2002, 90(3): 319−342

[66] Mastyło M. Bilinear interpolation theorems and applications[J]. Journal of Functional Analysis, 2013, 265(2): 185−207

[67] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//Proc of the 19th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 3354−3361

[68] Caesar H, Bankiti V, Lang A H, et al. nuScenes: A multi-modal dataset for autonomous driving[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 11621−11631

[69] Wu Xiaopei, Peng Liang, Yang Honghui, et al. Sparse fuse dense: Towards high quality 3D detection with depth completion[C]//Proc of the 39th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 5418−5427

[70] Yang Zeyu, Chen Jiaqi, Miao Zhenwei, et al. DeepInteraction: 3D object detection via modality interaction[C]//Advances in Neural Information Processing Systems. https://papers.nips.cc/paper_files/paper/2022/hash/0d18ab3b5fabfa6fe47c62e711af02f0-Abstract-Conference.html

[71] Song Ziying, Zhang Guoxing, Liu Lin, et al. RoboFusion: Towards robust multi-modal 3D object detection via SAM[J]. arXiv preprint, arXiv: 2401.03907, 2025

[72] Li Zhiqi, Wang Wenhai, Li Hongyang, et al. BEVFormer: Learning bird's-eye-view representation from LiDAR-camera via spatio-temporal transformers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(3): 2020−2036

[73] Liu Haisong, Teng Yao, Lu Tao, et al. SparseBEV: High-performance sparse 3D object detection from multi-camera videos[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 18580−18590

[74] Li Zhuoling, Xu Xiaogang, Lim S N, et al. Towards unified 3D object detection via algorithm and data unification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(9): 7960−7975

[75] Liu Yifan, Zhang Yong, Lan Rukai, et al. AWARDistill: Adaptive and robust 3D object detection in adverse conditions through knowledge distillation[J]. Expert Systems with Applications, 2025, 266: 126032

[76] Zheng Junhao, Lin Chenhao, Zhao Zhengyu, et al. Transferable stealthy attack methods for object detectors in autonomous driving systems[J]. Journal of Computer Research and Development, 2025, 62(8): 2111−2124 (in Chinese)
(郑君豪, 蔺琛皓, 赵正宇, 等. 针对自动驾驶系统目标检测器的迁移隐蔽攻击方法[J]. 计算机研究与发展, 2025, 62(8): 2111−2124)

[77] Hu Zixuan, Li Dongxiao, Ma Xinzhu, et al. Adaptive dual uncertainty optimization: Boosting monocular 3D object detection under test-time shifts[C]//Proc of the 20th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2025: 7273−7283

[78] Yoon J H, Jung J W, Yoo S B. Equirectangular point reconstruction for domain adaptive multi-modal 3D object detection in adverse weather conditions[C]//Proc of the 39th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025: 9553−9561

[79] Ibrahim M, Akhtar N, Wang H, et al. Multistream network for LiDAR and camera-based 3D object detection in outdoor scenes[J]. arXiv preprint, arXiv: 2507.19304, 2025



**Ji Wenyu**, born in 2002. Master candidate. His main research interests include deep learning and computer vision.

**纪文宇**, 2002 年生。硕士研究生。主要研究方向为深度学习、计算机视觉。



**Miao Zhuang**, born in 1976. PhD, professor. Member of CCF. His main research interests include artificial intelligence, pattern recognition, and computer vision.

**苗　壮**, 1976 年生。博士, 教授。CCF 会员。主要研究方向为人工智能、模式识别、计算机视觉。



**Zhang Qiyang**, born in 1998. PhD candidate. His main research interests include information security and object detection.

**张启阳**, 1998 年生。博士研究生。主要研究方向为信息安全、目标检测。



**Cui Xiaolu**, born in 2002. Master candidate. Her main research interests include computer vision and vehicle re-identification.

**崔小璐**, 2002 年生。硕士研究生。主要研究方向为计算机视觉、车辆重识别。



**Li Weixuan**, born in 2003. Master candidate. Her main research interests include artificial intelligence and computer vision.

**李唯暄**, 2003 年生。硕士研究生。主要研究方向为人工智能、计算机视觉。



**Li Yang**, born in 1984. PhD, associate professor. Senior member of CCF. His main research interests include computer vision, deep learning, and image retrieval.

**李　阳**, 1984 年生。博士，副教授。CCF 高级会员。主要研究方向为计算机视觉、深度学习、图像检索。