# 多关键点约束与深度估计辅助的单目3D目标检测算法

郑 锦[1),2)] 王 森[1)] 李 航[1)] 周裕海[1)]

[1)](北京航空航天大学计算机学院 北京 100191)

[2)](虚拟现实技术与系统全国重点实验室 北京 100191)

**摘 要** 当前主流的单目相机3D目标检测网络采用关键点检测范式,存在关键点预测与深度估计不准确的问题,限制了单目3D检测器的性能表现.本文提出一种多关键点约束与深度估计辅助的单目3D目标检测算法MonoAux,利用3D检测框的角点投影点、上表面与下表面中心投影点作为3D框中心投影点的补充,通过多关键点约束提升关键点预测精度;提出一种LiDAR-Free解耦深度估计方法,在不引入激光点云数据的同时通过几何关系推导引入额外的深度估计辅助监督信号,提升深度估计的准确性.多关键点约束与深度估计辅助仅在训练阶段使用,推理阶段不引入额外的计算成本.在KITTI3D目标检测验证集和测试集上的结果显示,相较于MonoDLE基线网络,提出的MonoAux算法在目标检测精度上分别提高3.87%和4.64%,与其他SOTA方法相比,本文方法也具有显著的性能优势,甚至优于部分使用额外数据的方法.

**关键词** 3D目标检测;关键点预测;角点投影点;深度估计;激光点云

**中图法分类号** TP391 **DOI号** 10.11897/SP.J.1016.2024.02803

# A Monocular 3D Object Detection Algorithm with Multi-Keypoint Constraints and Depth Estimation Assistance

ZHENG Jin[1),2)] WANG Sen[1)] LI Hang[1)] ZHOU Yu-Hai[1)]

[1)](*School of Computer Science and Engineering*, *Beihang University*, *Beijing* 100191)

[2)](*State Key Laboratory of Virtual Reality Technology and Systems*, *Beijing* 100191)

**Abstract** The mainstream monocular 3D object detection algorithms typically rely on a keypoint-based paradigm. While widely adopted, these approaches often face challenges in accurately predicting keypoints and estimating depth, which ultimately limit the performance of monocular 3D detectors. The core problem lies in the inherent difficulty of generating precise keypoints and depth values from a single 2D image. This paper introduces a novel solution to these issues, which is a monocular 3D detector named MonoAux that incorporates multi-keypoint constraints and depth estimation assistance. Traditional monocular 3D detection algorithms generally use the center projection point of the 3D bounding box as the primary keypoint for detection and localization tasks. However, relying solely on this center point often leads to suboptimal results, as it doesn't fully capture the spatial characteristics of the object. To improve the precision of keypoint prediction, MonoAux introduces multiple keypoints into the process. Specifically, it uses the corner points of the 3D bounding box and the center points of both the upper and lower surfaces of the bounding box. These additional keypoints serve as supplementary constraints to

收稿日期:2023-12-26;在线发布日期:2024-09-18. **郑 锦**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、目标检测与跟踪.E-mail:JinZheng@buaa.edu.cn. **王 森**,硕士,主要研究领域为计算机视觉、目标检测. **李 航**,博士研究生,主要研究领域为计算机视觉、目标检测. **周裕海**,硕士研究生,主要研究领域为计算机视觉、视频图像处理.

improve the prediction of keypoint prediction, and thus enhance the algorithm's ability to accurately estimate the object's orientation and shape in 3D space. By improving the prediction of these keypoints, MonoAux is able to generate more accurate 3D bounding boxes, which in turn improves the object detection performance. In addition to the multi-keypoint constraints, MonoAux introduces a novel approach to depth estimation that operates entirely without the use of LiDAR data. Many state-of-the-art (SOTA) 3D object detection methods rely on LiDAR point clouds to obtain accurate depth information, but this can be computationally expensive and requires specialized hardware. MonoAux tackles this challenge by proposing a LiDAR-free decoupling depth estimation method, which enhances the accuracy of depth estimation using only the geometric relationships inherent in the scene. This approach provides auxiliary supervision signals to improve the accuracy of depth prediction, even without the need for LiDAR data. As a result, the algorithm can estimate depth more accurately while maintaining efficiency and eliminating the need for expensive sensors. One of the key strengths of MonoAux is that the additional multi-keypoint constraints and depth estimation assistance are only applied during the training phase. This means that during the inference phase, there is no additional computational cost. The effectiveness of MonoAux is validated through experiments conducted on the KITTI3D object detection validation set and test set. These results show a substantial improvement in performance, with MonoAux achieving a 3.87% and 4.64% increase in object detection accuracy compared to the baseline network MonoDLE. Moreover, when compared to other state-of-the-art methods, MonoAux demonstrates significant performance advantages. It even outperforms some methods that rely on additional data, further proving its robustness and efficiency. In summary, MonoAux offers a significant advancement in monocular 3D object detection by addressing the core challenges of keypoint prediction and depth estimation. Its innovative use of multi-keypoint constraints and LiDAR-free depth estimation assistance not only improves accuracy but also ensures efficiency during the inference phase. The results on benchmark datasets underscore its potential to outperform existing methods, making it a promising solution for a range of applications.

**Keywords** 3D object detection; keypoint prediction; corner projection point; depth estimation; laser point cloud

## 1 引 言

3D目标检测能够提供物体中心点坐标、长宽高、偏向角等信息，是自动驾驶等任务不可或缺的能力[1]. 单目相机相较于双目相机具有结构简单、经济成本低、标定方便等优势，因此基于单目相机的3D目标检测算法研究得到越来越多的关注[2].

主流的单目3D目标检测网络如CenterNet[3]、MonoDLE[4]等，通常都基于关键点检测范式. CenterNet的核心思想在于将目标检测任务转换为先预测检测框中心点、再预测该点所代表物体的尺寸，进而预测出完整的检测框. CenterNet通过转换检测范式的方式，在检测速度与精度上均取得了显著的提升. MonoDLE是经典的端到端单目3D目标检测网络，设计思路继承于CenterNet，它不需要预先计算并设置大量锚框，加速了网络的推理效率. 这种基于关键点检测的网络相较于Anchor-based的单目3D检测器普遍拥有更强的性能优势.

然而，对MonoDLE深入分析发现，造成其检测精度依然不高的最根本原因在于仅利用3D框中心投影点得到的关键点预测定位不准、深度估计有待改进. 因此，本文提出单目3D目标检测算法MonoAux(Monocular 3D object detection Auxiliary)，引入辅助学习的思想，设计多关键点约束与深度估计辅助，在训练阶段帮助检测头训练. 主要贡献如下：

(1)提出3D检测框投影点辅助学习模块，利用

3D检测框的角点投影点、上表面与下表面中心投影点作为3D框中心投影点检测头训练的补充，通过多关键点约束提升关键点预测精度，为投影点预测提供更多监督信号；

(2)提出LiDAR-Free解耦深度估计算法，在不引入额外的激光雷达数据、也无须依赖深度补全算法的情况下，仅通过几何关系推导获取深度估计辅助监督信号，提升深度估计准确性；

(3)在KITTI3D数据集上的实验表明，提出的MonoAux集成了多关键点约束与深度估计辅助，相较于基线模型取得了显著的性能提升，在KITTI3D验证集和测试集上，MonoAux相对基线MonoDLE，目标检测精度分别提高3.87%和4.64%. MonoAux算法无需依赖激光雷达，同时不引入额外的推理计算量，可被应用于任何基于关键点检测范式的单目3D目标检测器，实现在推理速度不变情况下的检测性能提升.

## 2　相关工作

### 2.1　单目3D目标检测网络基本框架及问题分析

在基于关键点检测范式的单目3D目标检测网络中，输入图片经过特征提取、特征融合、独立检测头三部分，预测输出目标的3D检测框. 检测结果由定位$(x,y,z)$、三维尺寸$(w,h,l)$和偏航角$(\theta)$表示.

在特征提取部分，基于关键点检测的网络主要选择DLA[5]、Hourglass[6]作为骨干模型进行特征提取. 随后，与成熟的2D目标检测类似[7]，单目3D目标检测中也继承了Neck层，Neck层对输入的骨干特征依次进行“上采样-聚合”变换后，将多尺度特征进行融合，兼顾了深层强语义与浅层高精度定位的优势. 进而，检测头部分被设计为多个独立的预测分支，包括关键点分类头、深度估计头、3D偏移量预测头、3D检测框尺寸头、偏航角预测头、2D检测框、2D偏移量预测头，最后预测输出目标的3D检测框.

作为经典的基于关键点检测范式的单目3D目标检测网络，MonoDLE工作中着重分析了目前单目3D目标检测的性能瓶颈，本文将MonoDLE结论总结如下表1所示.

**表1　单目3D目标检测网络消融实验结果[4]**

| 基线模型 | ＋标注投影点 | ＋标注深度 | ＋标注3D定位 | ＋标注3D尺寸 | ＋标注偏航角 |
|---|---|---|---|---|---|
| 11.12 | 23.90 | 38.01 | 78.84 | 11.96 | 11.88 |

在表1中，第1列表示基线模型CenterNet的检测性能，mAP仅为11.12%，第2～6列表示将基线模型对应检测头的结果替换为数据集相应标注之后的检测性能. 由表1可知，基线模型在对3D检测框尺寸的预测以及朝向的学习上已经达到了令人满意的结果，即使用标注真值替换基线方法中的相应预测值，依然不能带来明显的检测精度提升；但在替换3D定位、深度标注真值后，性能得到大幅提升. 可见，在3D关键点定位以及深度估计上仍然存在巨大的提升空间. 图1的结果也验证了这一点，其中，绿色框为标注框，红色框为MonoDLE预测框，显然，两者在尺寸、朝向上差异并不明显，而定位和深度存在明显区别.

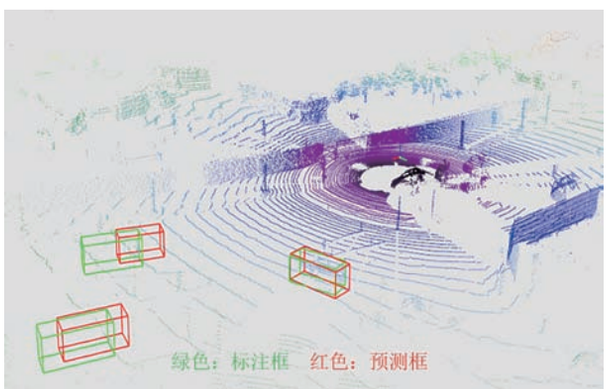



图1　MonoDLE网络3D目标检测结果与标注真值

### 2.2　单目3D目标检测中的关键点检测

虽然目前KITTI榜单上排名靠前的单目相机3D目标检测方案均采用关键点检测范式，但其性能依然远低于基于激光点云的3D检测器. 上述表1和图1的结果明确指出其检测精度依然不高的主要原因之一在于定位误差过大. 事实上，仅通过3D检测框中心投影点训练网络，用单点表示整个物体会存在关键点定位不准、进而影响3D目标检测精度的问题.

针对MonoDLE相关实验揭露的关键点定位不准的问题，MonoCon[8]在MonoDLE的基础上额外引入检测框的8个角点开展辅助训练，涉及到的辅助训练分支包含关键点预测、偏移量分支以及中心点-角点偏移量预测分支，在MonoDLE的基础上显著提升了检测性能. 然而，MonoCon参与辅助训练的这8个角点是各自单独训练的，预测框的整体结构考虑不足，仍然存在角点单独训练从而得到次优解的问题. 此外，MonoCon未能在深度估计分支提出改进措施. AutoShape[9]认为现有的基于关键点的

检测器并不能很好地利用目标的形状信息，因此设计了一种自动标注算法以生成带有形状信息的多个关键点标注，基于更多的关键点训练以及预测取得了一定的性能提升，但复杂度较高．现有基于关键点的单目3D目标检测算法往往集成多个关键点预测角点深度，期望利用多个关键点的特征以提高算法的准确率和鲁棒性，但这类算法中关键点的特征往往相似，集成时不能达到纠正偏差的目的，为此，MonoCD[10]结合全局特征与局部特征，同时利用关键点间的几何关系来实现深度预测数值上的互补，提高模型性能．

3D目标拥有丰富的潜在关键点信息，在中心投影点的基础上加入更多角点，甚至是带有形状信息的多个关键点标注是一种提升3D关键点定位准确性的方法．如何选择关键点位置及个数，如何利用关键点之间的相对位置关系，在保证精度的同时控制计算复杂度是值得研究的问题．

### 2.3 单目3D目标检测中的深度估计

MonoDLE认为目前深度模型几乎不可能获得完全准确的深度估计结果，并提出了一种样本过滤操作，即过滤超过60 m的训练样本，从而让检测器更加关注60 m以内样本的检测效果．而更多的研究者依然在不断尝试如何提升单目3D目标检测深度估计的准确性．

MonoRCNN[11]是一种基于区域的双阶段单目3D目标检测器，它提出一种基于几何分解的深度估计策略，通过将深度估计分解为物体物理高度与图像平面投影高度，使得深度估计更加具有可解释性，并提升了深度估计的稳定性与准确性．GUPNet[12]是近两年应用最为广泛的基于区域的双阶段单目3D检测器，GUPNet与MonoRCNN同样在引入几何高度先验的同时加入深度估计不确定性．然而，与MonoRCNN不同的是，GUPNet以预测的投影高度为基础，并将不确定性通过预测实际高度传递到最终计算深度，而MonRCNN则以预测实际高度为基准并将不确定性通过预测投影高度传递到深度计算中．两者均尝试利用几何投影模型得到较为准确的深度预测值，同时两者均通过不确定性对深度预测进行修正．然而，针对远距离物体的投影高度预测往往不够准确，从而导致最终深度估计的误差较大．DID-M3D[13]是基于GUPNet改进的一种解耦深度单目3D检测器．与GUPNet一样，DID-M3D首先预测2D检测框并根据2D检测框区域特征预测3D检测框．DID-M3D提出将目标的实例深度解耦为可视深度以及属性深度，得益于属性深度预测更加容易，因此解耦深度的策略能够进一步缓解实例深度估计误差大的问题．2021年提出的MonoFlex[14]在直接预测深度的基础上引入几何投影模型的深度先验知识，通过集成多个预测的深度值，结合提出的截断物体检测模块，MonoFlex取得了当时最先进的性能表现．

此外，在MonoRCNN基础上，MonoRCNN++[15]将物体的真实高度和视觉投影高度进行联合建模，帮助模型显式地学习两种高度之间的关联，同时使用不确定感知回归损失替代原方法中的L1损失，来提高模型对物理尺寸、偏航角和投影中心的预测精度．MonoDETR[16]将DETR框架用于单目3D目标检测，在原框架中添加深度编码器，并使得object query与depth token完成特征交互，使得object query中包含深度信息，并在此基础上捕获几何线索，在利用有限视觉几何特征的同时将信息丰富的深度特征加入到模型中训练．MonoDTR[17]提出深度特征增强模块，在不增加计算成本的同时通过辅助监督信号隐式学习深度特征．BEVDepth[18]提出深度细化模块，将深度监督信号引入到模型训练中，一定程度上缓解了特征与位置不对应的问题，提高了模型的检测性能．NeRF-Det[19]指出现有3D目标检测算法不能很好地编码场景的几何信息，提出将NeRF引入检测模型中，使检测模型与NeRF共用一个MLP，同时用深度损失来监督NeRF渲染出深度图，以此约束检测模型编码场景的几何信息和深度信息．

另一方面，目前主流的单目3D目标检测器（如MonoDLE、GUPNet等）都是直接预测3D目标检测框的中心点深度，这种用单个关键点深度代表整个物体深度的方式会导致潜在的训练不足．为了进一步增加对深度估计预测头的监督强度，一种做法是引入额外的激光雷达数据作为深度监督信号；另一种做法是直接将感兴趣物体区域的深度统一用物体的3D标注深度表示，从而形成一个粗粒度的深度监督信号．然而这两种方式都存在一定的缺陷，前者采用激光雷达信号作为输入，保证在深度估计结果的精度上具有优势，但引入了额外数据，而且激光雷达对透明车窗等物体的深度估计不够准确；后者虽然摆脱了额外数据的引入以及透明物体估计不准的问题，但是存在精度不够细致的问题．如果能够在不引入激光点云数据的同时，仅通过简单的几何关系推导就能获取目标各个区域较为准确的深度估计

结果，并将其作为辅助监督信号进行网络训练，则能有效提升检测性能.

综上所述，基于关键点检测范式的单目3D检测器通过3D检测框中心投影点训练网络，用单点表示整个物体会存在关键点定位不准以及深度估计训练不充分的问题. 本文以此为出发点，显式引入更多关键点辅助学习以加强目标结构，进而提升网络整体关键点的预测精度，并通过几何关系推导引入更多的深度监督信号以提升检测器整体的深度估计准确性，最终提升网络的检测性能.

## 3　本文方法

针对主流单目3D目标检测器部分检测头学习不充分的问题，本文提出关键点约束与深度估计辅助的单目3D目标检测算法MonoAux，设计3D检测框投影点辅助学习模块与LiDAR-Free深度估计辅助学习模块，提升3D目标检测精度. 这两个模块仅在训练阶段使用，推理阶段无需使用，因此不会增加网络推理负担.

### 3.1　MonoAux网络总体框架

MonoDLE作为经典的单目3D目标检测网络，无论在精度上还是速度上均有明显优势，因此本文将其作为基础3D检测器. 图2为本文提出的MonoAux网络总体框架.

首先，KITTI数据集作为网络的输入送入骨干网络DLA34中，产生的四层特征图经过Keypoint特征融合层并产生相较原图四倍下采样的特征图. 在获取网络中FPN层的特征图后，该特征图被送到多个独立检测头中，将这些独立检测头按照功能划分到以下分支：3D检测分支、2D检测分支、投影点辅助训练分支、深度估计辅助训练分支. 前两个分支为MonoDLE已有分支，投影点辅助训练分支、深度估计辅助训练分支是MonoAux新增分支.

投影点辅助训练分支直接基于特征融合层的输出特征图开展预测，这一模块内三个检测头设计与2D检测分支、3D检测分支的检测头完全相同，均采用“卷积-ReLU-卷积”的组合方式，唯一不同的是最后一层卷积的输出通道数不同. 深度估计辅助学习分支利用2D检测分支预测的2D检测框抽取出目标区域的特征，经过ROI-Align处理后统一缩放至7×7大小. 同样的，采用“卷积-ReLU-卷积”的组合方式设计表面深度估计头与偏移深度估计头，表面深度估计头与偏移深度估计头分别预测表面深度和偏移深度，直接与伪深度图计算损失. 这里，伪深度图作为真值进行训练，即图2中的LiDAR-Free深度标签. 伪深度图中的表面深度根据提出的LiDAR-Free解耦深度估计算法由几何规则推导出来，偏移深度是3D目标检测中心点与表面深度的差值.

在训练阶段，图2中所有独立检测头均参与训练；而在测试推理阶段，投影点辅助训练分支、深度估计辅助训练分支分别对应的3D检测框投影点辅助学习模块、LiDAR-Free深度估计辅助学习模块并不需要参与，因此MonoAux能够在不引入额外推

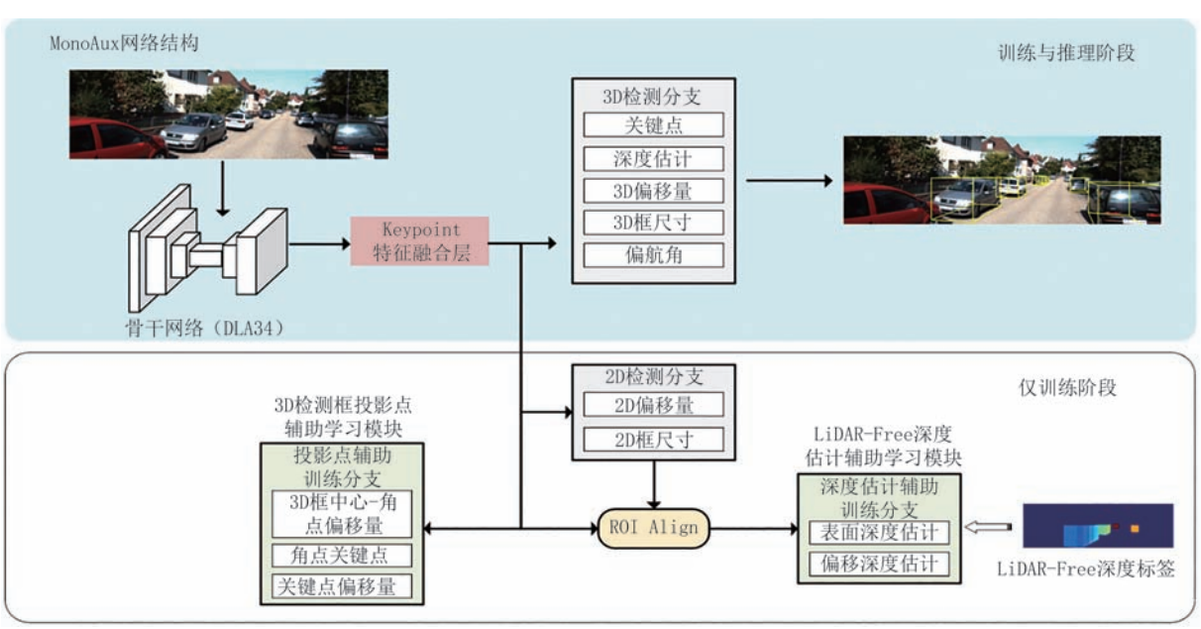



图2　基于辅助学习的单目3D检测网络MonoAux

理计算量的情况下，极大提升单目3D检测器的计算效率和检测精度.

### 3.2 3D检测框投影点辅助学习模块

主流的3D检测器在预测3D检测框时将问题分解为预测3D检测框的中心点、三维尺寸以及朝向.考虑到输入图像是二维平面，通常做法是将3D检测框的中心点投影至二维图像平面并预测，而丢失的深度信息则利用密集深度估计尝试恢复.投影点预测不准确的问题实际上可以归纳为3D检测框中心点的投影点预测分支的精度不够高.为了给投影点预测分支提供更多监督信号，同时兼顾3D目标检测任务的特性，本文利用更多投影点开展训练，通过设置额外独立的关键点预测分支，保证中心点预测和其他关键点预测目标不冲突，通过多关键约束提升关键点预测的准确性.

#### 3.2.1 角点关键点预测

单目图像中的3D框中心投影点以及3D检测框的8个角点是一组适定的投影二维监督信号.MonoCon[8]率先提出使用3D检测框的角点投影点作为3D框中心投影点检测头训练的补充，本文同样采用了这种方式.为了让角点投影点预测结果更加符合3D检测框的几何形状，本文进一步在图像平面内同步引入额外两个面的中心投影点，即3D检测框的上表面与下表面的中心点，以加强对角点投影点预测的约束.这10个投影点的预测统称为角点关键点预测.其中，新增的两个表面角点是基于预测的8个角点再计算出来进行训练的，考虑了预测框的整体结构，即新增的两个表面角点根据上下表面中心点的中点即为3D检测框中心点这一事实设计损失函数，从而迫使投影点预测模块在训练过程中更加关注整体形状.虽然表面看相比MonoCon只是增加了2个角点，但是角点之间形成了位置约束关系，新增的上下表面中心点的中点即为3D检测框中心点，能够迫使模型得到更加准确的整体形状，适应性更强.

在3D检测框投影点辅助训练分支中，角点关键点检测头中采用的损失函数与中心投影点的损失函数一致.具体而言，考虑到关键点检测范式的网络中物体区域只有标注投影点的像素点才归为正样本，而其他像素均被设为负样本，因此这种训练方式面临极为严重的正负样本不平衡问题.为了克服目标定位的正负样本不平衡问题，本文采用Gaussian Focal Loss.

假设$x_b$与$y_b$为2D检测框中心点在特征图上的图像坐标，则对中心点$(x_b,y_b)$建模高斯核$G$的计算过程如下：

$$G(x,y)=\exp\left(-\frac{(x-x_b)^2+(y-y_b)^2}{2\times\sigma_b^2}\right) \tag{1}$$

其中$(x,y)$表示高斯核预先设定的区域内的坐标，$\sigma_b$表示预定义的对象大小自适应标准偏差.图3可视化了热力图(Heatmap)与RGB图像叠加的结果，即在10个待预测角点投影点处叠加10个Heatmap.在这个过程中，如果两个Heatmap重叠，则重叠区域取最大值.在预处理角点投影点Heatmap之后，Gaussian Focal Loss计算过程如下式所示：

$$L(H,H^*)=-\frac{1}{N}\sum_{(x,y)}\begin{cases}(1-H_{xy})^{\gamma}\log(H_{xy}), & \text{if } H^*_{xy}=1\\(1-H^*_{xy})^{\beta}(H_{xy})^{\gamma}\log(1-H_{xy}), & \text{Otherwise}\end{cases} \tag{2}$$

其中，$H$与$H^*$分别表示模型预测的Heatmap以及真实Heatmap，$N$表示真实物体的个数，$\gamma$与$\beta$则是超参(通常分别设置为2.0与4.0).

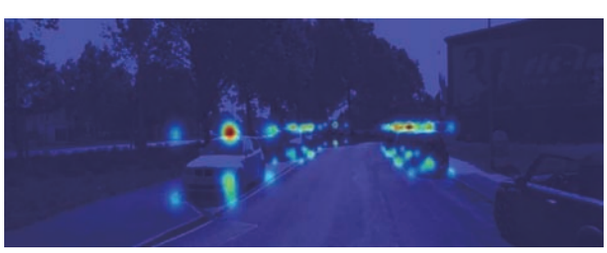


图3 每个目标上10个关键点热力图

#### 3.2.2 关键点偏移量预测

考虑到用于预测的特征图相较于原图输入往往存在降采样问题，直接在特征图中预测准确目标位置存在误差，因此提出关键点偏移量预测头.假设在原图中目标中心点为$(x^*_b, y^*_b)$，当前特征图的目标坐标为$(x_b, y_b)$，则$(x_b, y_b)$与$(x^*_b, y^*_b)$坐标之间的关系计算如下：

$$\begin{cases}x_b=\left\lfloor\dfrac{x^*_b}{s}\right\rfloor\\y_b=\left\lfloor\dfrac{y^*_b}{s}\right\rfloor\end{cases} \tag{3}$$

其中$s$表示原图与预测特征图之间存在$s$倍下采样倍率.关键点偏移量预测头计算如下：

$$L_{ht\_offset}=F1(H_{kpt},H^*_{kpt}) \tag{4}$$

其中，$L_{ht_offset}$表示关键点降采样误差的损失函数，而$H_{kpt}$与$H^*_{kpt}$分别代表模型预测的关键点偏移量以及

真实关键点偏移量，计算采用简单的回归损失函数$F1$.

3.2.3　中心投影点与角点投影点偏移量预测

除了显式地让模型预测新增的10个角点投影点位置，本文还引入3D框中心点投影点与3D框角点投影点之间的偏移量预测子任务．增加对3D框中心投影点与角点投影点之间的偏移量预测，一方面可以间接为其他关键点偏移量预测分支的训练提供额外监督信号，另一方面该任务搭配角点投影点预测任务，可以为角点投影点预测提供更多线索．如图4所示，红色点、蓝色点分别表示3D检测框中心点、10个待预测角点投影点，黄色虚线箭头表示从中心点到角点投影点的偏移量．通过在2D图像平面内让模型显式地预测中心投影点与角点投影点之间的偏移量，能够进一步提升模型对角点投影点的预测精度．与MonoDLE基线模型自有的偏移量预测头一样，这里所使用的偏移量预测头基于浮点数预测，而并非像素平面内的整型数，因此可以避免训练过程中的精度丢失．

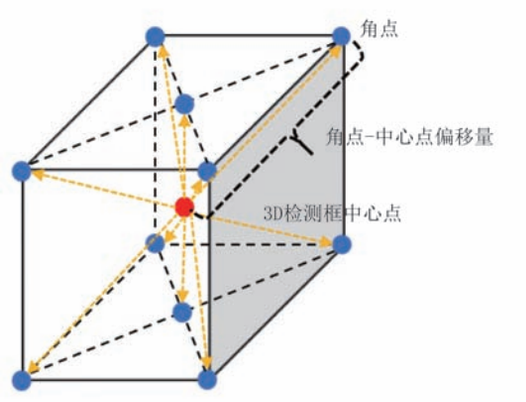



图4　中心投影点与角点投影点的偏移量预测

为监督训练这一分支，采用回归损失函数，其计算过程如下：

$$L_{kpt\_offset}=F1\left(o_{kpt},o^{*}_{kpt}\right) \tag{5}$$

其中，$L_{kpt_offset}$表示该中心投影点与角点投影点偏移量的损失函数，而$o_{kpt}$与$o^{*}_{kpt}$分别代表模型预测的偏移量以及真实偏移量．为了简化问题，该分支直接采用$F1$损失函数计算．

### 3.3　LiDAR-Free深度估计辅助学习模块

本文设计LiDAR-Free深度估计辅助学习模块，能够在不引入额外的激光雷达数据、不依赖深度补全算法的情况下，仅通过几何关系推导就获取感兴趣物体区域的深度估计结果．此外，该深度估计辅助学习模块能嵌入任何联合深度估计训练的网络，并稳定带来明显的性能增益．

3.3.1　解耦深度估计策略分析

通过将检测框的中心深度分解为表面深度与偏移深度，模型的单点深度估计任务可以进一步转化为预测可视表面上的多点深度，由此引入更多的深度估计监督．具体而言，图5中黄线表示BEV视角下的表面深度，绿线表示当前视角下的可视物体表面，而红线表示不可视深度．可视物体表面的判断以及表面深度的计算均可根据几何规则计算．DID-M3D[13]也提出解耦深度估计的思想，但本文提出的方案与DID-M3D完全不同：(1)DID-M3D将解耦深度直接嵌入单目3D检测器，并在推理阶段继续使用这一解耦的多点深度估计模块，而本文的解耦深度仅用于辅助训练，能够以一种更加经济的方式提升网络性能；(2)DID-M3D依赖激光雷达输入数据，通过使用离线的深度补全算法预先处理得到稠密深度图，以获得精确的单目深度图，而本文方案可以根据潜在的几何规则直接计算深度图，从而摆脱对额外激光点云数据以及点云深度补全算法的依赖．因此，本文提出的LiDAR-Free深度解耦辅助训练策略更加灵活．

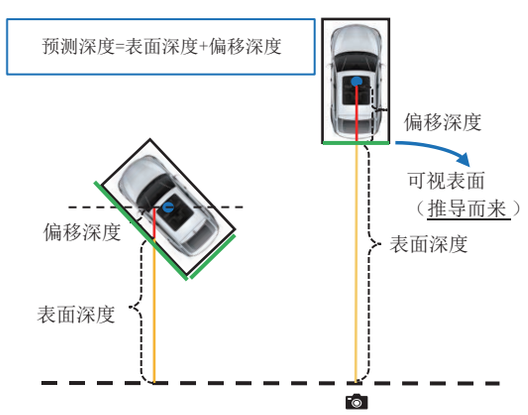



图5　MonoAux解耦深度估计示例图

3.3.2　LiDAR-Free解耦深度估计算法

本文提出一种根据3D标注框、基于几何推导直接生成目标深度图的方法，即LiDAR-Free解耦深度估计算法．如图6所示，具体流程如下：

(1)3D检测框转BEV坐标：在已知3D检测框的标注信息之后，计算出3D检测框的八个角点在三维空间内的坐标，取上表面或下表面四个顶点，去除高度信息后，将三维坐标转化为二维BEV坐标．

(2)确定可视区域左右两端角点：根据相机成

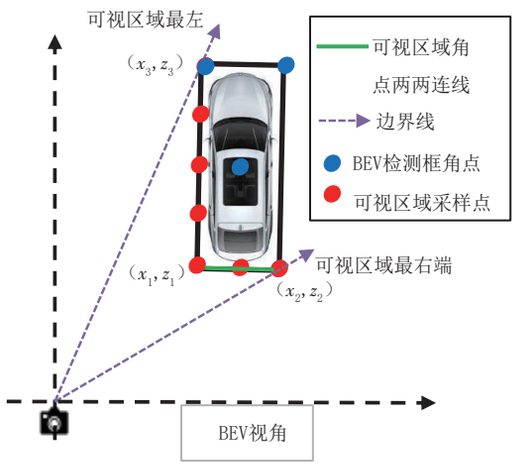



图6 以目标为中心的LiDAR-Free深度图生成示例图

像特性确定相机能够看到的最左顶点$(x_3,z_3)$与最右顶点$(x_2,z_2)$.

(3)确定其他可视顶点(如$(x_1,z_1)$):已知两点形成直线($x_1=x_2$特殊处理),存在关系$(x-x_1)/(x_2-x_1)=(z-z_1)/(z_2-z_1)$,因此,给定$x$,深度$z$计算公式如下:

$$z=\frac{(z_2-z_1)x+(x_2z_1-x_1z_2)}{x_2-x_1} \tag{6}$$

其中,$z$是表面深度,$(x,z)$为可视区域采样点坐标.

(4)根据预设采样密度在三维空间采样:在获得可视角点集合后,根据直线深度均匀变化规则,确定采样的$x$坐标密度,即体素(Voxel)大小,依次计算每一采样点表面深度$z$. 偏移深度是3D目标检测中心点与表面深度的距离,用3D目标检测中心点与表面深度的差值表示,偏移深度可以是负数.

(5)根据采样三维点投影回2D平面构建深度图:根据采样点投影回2D图像平面内,最终生成以目标为中心的深度图.

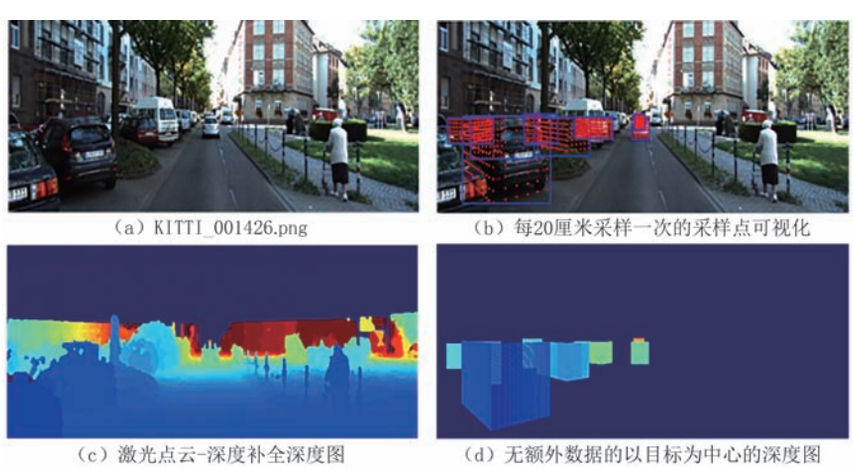
图7(a)为KITTI3D数据集原图,图7(c)展示了基于激光点云投影生成的稠密深度图,图7(d)展示了本文LiDAR-Free解耦深度估计算法生成的深度图,可以观察到该图较为精准地刻画了物体与物体之间的深度关系. 图7(b)则展示了在三维空间内每隔20 cm采样一次,在图像平面内对应的像素点. 考虑到三维数据投影回二维平面中并不能保证均匀采样的特性,因此图7(b)中对应采样点并非均匀分布的,也从侧面验证了本文LiDAR-Free算法的正确性. 本文提出的MonoAux基于LiDAR-Free生成的伪深度图开展训练.

## 4 实验分析比较

### 4.1 数据集及实验设置

本文采用KITTI3D数据集进行实验,该数据集包含了14999张3D目标检测双目图像,针对单目3D目标检测任务一般仅采用左图作为模型输入. 训练数据包括7481张带标注图像,测试集包含7518张未公开标注的图像用于官方测评. 实验采用VoxelNet[20]中所采用的划分方式对标注图像切分,将7481张训练数据切分为3712张训练集以及3769张验证集,保证性能指标对比的公平性.

实验中所有消融实验采用相同的超参数,选择Adam[21]优化器,其参数设置$\beta_1=0.9$,$\beta_2=0.999$;学习率调度器选择LambdaLR,训练前5轮默认允许采用Warm up热身策略[22]. 消融实验中设置训练的Batch size为16,一共训练160轮并每隔10轮保留模型,消融实验中展示所有中间模型的最佳性能结果. 在实验第5至95轮采用0.001 25学习率,第



(a) KITTI_001426.png (b) 每20厘米采样一次的采样点可视化

(c) 激光点云-深度补全深度图 (d) 无额外数据的以目标为中心的深度图

图7 提出的LiDAR-Free深度图效果图

96至125轮采用0.000 125学习率，最后35轮采用0.000 0125的学习率训练. KITTI3D的彩色图片在训练前统一缩放到1280×384分辨率，同时训练过程中仅能使用随机翻转、随机裁剪两种数据增强手段.

实验中采用评价指标为3D AP、BEV AP以及AOS. 其中Easy、Mod、Hard分别表示简单目标、一般目标、困难目标. 所有指标均越高性能越好.

实验采用Intel(R) Xeon(R) Gold 6248 CPU @ 2.50GHz，GPU为NVIDIA GeForce 3090，内存为512GB，操作系统为Ubuntu 18.04，训练框架为Pytorch 1.12.0+cu113.

### 4.2　3D检测框投影点辅助学习模块实验

本文首先基于KITTI3D验证集开展消融实验，验证投影点辅助学习模块的有效性. 结果见表2.

由实验结果可以看出，本文设计的投影点辅助学习模块稳定提升了单目3D检测器MonoDLE的性能. 表2第二行代表本文复现的MonoDLE精度，可以看到本文复现的精度与原论文精度相近. 当引入本文使用的8个角点投影点预测后，3D AP指标相较于复现精度分别取得了2.6%、1.69%和2.07%的绝对性能提升(19.53vs. 16.93、15.76vs. 14.07、14.06vs. 11.99)，证明了引入投影点辅助学习子任务能够有效带来性能提升. 值得注意的是，仅引入中心投影点与角点投影点偏移量预测子任务后，模型并未带来明显的性能提升，甚至在某些指标下下降，我们发现原因在于其损失较大，该训练分支难收敛，存在学习困难的问题. 事实上，中心投影点与角点投影点偏移量的估计并不是我们的目的，我们的最终目标还是要确定目标框的位置，也就是角点预测是更为重要的. 中心投影点与角点投影点偏移量可为角点投影点预测提供更多线索，引入投影点预测，可以有效降低模型对中心投影点与角点投影点偏移量预测的学习难度，同时这两个分支可以相互辅助并带来更精准的关键点预测定位性能，从而带来整体的性能提升. 为了验证中心投影点与角点投影点偏移量分支是否学习充分，通过将上述两辅助分支统一联合训练之后，表1的第五行结果显示联合训练两个子任务可以带来性能提升，3D AP绝对性能分别提升3.09%、1.76%和2.12%(20.02vs. 16.93、15.83vs. 14.07、14.11vs. 11.99). 除此之外，引入上表面与下表面的中心投影点之后，这两个投影点分别对上表面以及下表面8个关键点起到约束作用，从而形成更加符合立方体形状的预测. 表2最后一行的结果显示引入这两个关键点能够带来稳定的性能提升，相比8个角点，10个角点在3D AP指标上分别带来0.29%、0.69%和0.18%的绝对性能提升(20.31vs. 20.02、16.52vs. 15.83、14.29vs. 14.11). 总的来看，本文提出的投影点辅助学习模块相比基线方法MonoDLE，在3D AP指标上分别带来3.38%、2.45%、2.3%绝对性能提升(20.31vs. 16.93、16.52vs. 14.07、14.29vs. 11.99). 由此可见，本文提出的投影点辅助学习模块是有效的，通过多关键点约束提升了3D检测精度.

表2同时从推理时间、参数量这两个方面进行了模型复杂度分析. 实验结果可以看出，本文设计的投影点辅助学习模块稳定提升了单目3D检测器MonoDLE的性能，但是并不会引入过多参数量；投影点辅助训练分支在推理阶段并不会参与运行，因此本文提出的结合了角点投影点预测、中心投影点与角点投影点偏移量的模型，其推理时间和MonoDLE保持一致.

### 4.3　LiDAR-Free深度估计辅助学习模块实验

为了验证LiDAR-Free深度估计辅助学习模块，本节选择表2中集成了10个角点投影点预测、10个

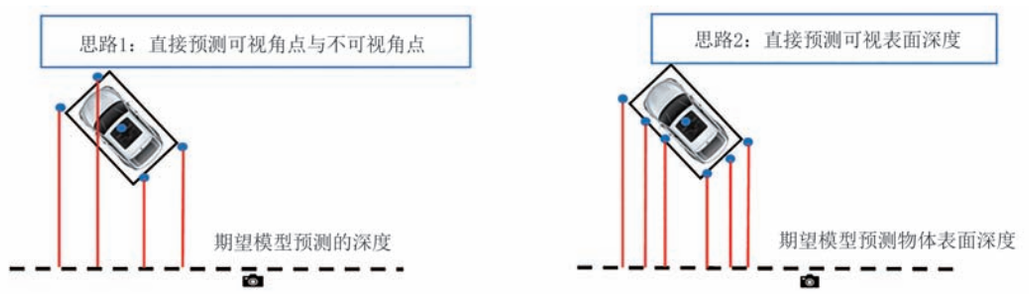
**表2　投影点辅助学习模块在KITTI3D验证集上的结果**

| IoU阈值(0.7) | 3D AP ↑ | | | BEV AP ↑ | | | AOS ↑ | | | 推理时间 | 参数量 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Easy | Mod | Hard | Easy | Mod | Hard | Easy | Mod | Hard | (ms) | (M) |
| MonoDLE论文结果[4] | 17.45 | 13.66 | 11.68 | 24.97 | 19.33 | 17.01 | - | - | - | - | - |
| MonoDLE复现结果 | 16.93 | 14.07 | 11.99 | 23.58 | 19.34 | 17.49 | 97.60 | 90.93 | 82.69 | 40 | 20.32 |
| MonoDLE+8个中心投影点与角点投影点偏移量 | 16.47 | 14.07 | 12.03 | 23.38 | 19.36 | 17.60 | 97.54 | 91.39 | 81.36 | 40 | 20.46 |
| MonoDLE+ 8个角点投影点预测 | 19.53 | 15.76 | 14.06 | 27.03 | 22.08 | 19.28 | 97.79 | 91.52 | 83.35 | 40 | 20.61 |
| MonoDLE+8个角点投影点预测+8个中心投影点与角点投影点偏移量 | 20.02 | 15.83 | 14.11 | 28.24 | 22.26 | 19.50 | 97.73 | 91.63 | 83.73 | 40 | 20.69 |
| MonoDLE+10个角点投影点预测+10个中心投影点与角点投影点偏移量(本文方法) | 20.31 | 16.52 | 14.29 | 28.62 | 22.37 | 19.58 | 98.24 | 92.05 | 84.21 | 40 | 20.76 |

中心投影点与角点投影点偏移量的模型作为基线模型，增加多种不同深度估计策略，验证其性能.

针对深度估计，一种直观的想法就是直接迫使检测器去预测出3D检测框8个角点的深度（如图8(a)），因为3D目标检测框在三维世界的坐标是已知的. 然而，表3第2行的实验结果揭示了这种方案是无效的，这是因为预测3D框中心点深度的任务与预测8个角点的深度是完全不相同的预测目标，同时，直接让网络预测物体背面不可见关键点的深度是极为困难的，在训练实验过程中也观察到损失函数无法收敛.

此外，还有一种做法是直接预测表面深度（如图8(b)），同样地，预测3D检测框的中心点与预测物体表面深度是两种不同目标的预测任务，表3第3行的结果也显示了这种方案并不能得到最好的结果，模型在绝大多数指标上仅取得了轻微的性能提升.




（a）直接预测角点深度　　（b）直接预测可视深度

图8　错误的深度预测思路

**表3　不同深度辅助策略在KITTI3D验证集上的结果**

| IoU阈值 | 3D AP ↑ | | | BEV AP ↑ | | | AOS ↑ | | | 推理时间 | 参数量 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (0.7) | Easy | Mod | Hard | Easy | Mod | Hard | Easy | Mod | Hard | (ms) | (M) |
| MonoDLE+10个角点投影点预测+10个中心投影点与角点投影点偏移量（基线模型） | 20.31 | 16.52 | 14.29 | 27.53 | 22.37 | 19.58 | 98.24 | 92.05 | 84.21 | 40 | 20.76 |
| +8个角点深度直接预测（图8(a)） | 17.69 | 14.62 | 12.56 | 24.22 | 19.92 | 18.31 | 98.04 | 89.59 | 83.75 | 40 | 20.83 |
| +表面深度直接预测（图8(b)） | 19.56 | 16.55 | 14.42 | 27.70 | 22.42 | 20.61 | 96.08 | 92.21 | 84.30 | 40 | 20.86 |
| +以标注深度为表面深度 | 20.68 | 16.04 | 14.35 | 29.68 | 22.73 | 19.87 | 97.87 | 91.61 | 81.55 | 40 | 20.76 |
| +IP_basic[23]深度图 | 20.84 | 17.07 | 14.68 | 28.62 | 22.91 | 20.84 | 98.86 | 92.81 | 84.71 | 40 | 20.76 |
| +PENet[24]深度图 | 22.18 | 17.80 | 15.34 | 30.24 | 23.45 | 21.36 | 96.02 | 92.31 | 84.14 | 40 | 20.76 |
| +LiDAR-Free深度图（本文方法） | 21.60 | 17.53 | 14.94 | 29.48 | 24.12 | 21.16 | 95.86 | 92.04 | 83.95 | 40 | 20.76 |

而若将目标区域的像素深度统一用3D框标注深度代替（表3第4行），检测性能在大部分指标上取得提升，这是因为此时与直接预测中心点深度具有相同的回归目标，但性能提升不够明显.

考虑到上述几种深度估计辅助学习策略提升有限，深度估计任务需要被解耦. 在保证回归目标是预测3D框中心点深度的同时，还要保证网络只能预测物体可视表面，实现解耦的深度估计. 表3最后三行所展示的实验结果均采用解耦深度估计的思路，唯一不同的是在辅助深度估计学习模块中采用的深度图不同. 其中，IP_Basic[23]是一种不可学习的深度补全算法，而PENet[24]是一种基于深度学习的深度补全算法，它们均依赖激光点云数据.

当对稀疏深度图采用IP_Basic深度补全算法后，3D检测器的性能得到了明显提升，其中AOS指标在表3中取得了最佳表现. 当使用PENet深度补全算法时，模型在绝大多数AP指标上取得了最好表现. 然而，无论是IP_Basic还是PENet均依赖于激光点云数据，而本文LiDAR-Free深度图则直接根据几何规则推导计算得到. 尽管不使用激光雷达数据，本文LiDAR-Free深度图生成方案整体上依旧优于使用激光点云的IP_Basic深度补全方案，而LiDAR-Free深度图训练的单目3D目标检测器在性能上也仅稍弱于PENet. 因此，表3中伪深度图与激光雷达深度图用于3D目标检测时性能相近可以佐证本文深度估计结果的准确性，进一步凸显LiDAR-Free深度图拥有与激光点云深度图相近的性能. 通过详尽的对比实验，本文提出的LiDAR-Free深度估计辅助训练模块的有效性得到验证. 综

合来看，集成了多关键点约束与LiDAR-Free深度图的本文方法，相比MonoDLE在3D AP Mod指标上提升了3.87%（17.53vs. 13.66）.

此外，表3同时从推理时间、参数量这两个方面进行了模型复杂度分析. 实验结果可以看出，本文LiDAR-Free深度估计辅助学习得到的LiDAR-Free深度图有效支持了3D目标检测，在稳定提升单目3D检测器MonoDLE的性能的同时，并不会引入更多的参数量；深度估计辅助训练分支在推理阶段并不会参与运行，因此本文提出的结合了LiDAR-Free深度估计辅助学习模块的模型，其推理时间和MonoDLE保持一致.

此外，本文进一步和典型的基于深度估计的单目3D目标检测方法MonoGRNet、D$^4$LCN、MonoPSR等进行了比较，这三种方法和MonoDLE、MonoDLE＋LiDAR-Free深度图（本文方法）的检测结果见表4. 从表中可以得出，我们的方法明显优于除了D$^4$LCN之外所有方法，与D$^4$LCN检测结果相比，即使仅在MonoDLE基础上增加LiDAR-Free深度图，检测效果也相当，证明了LiDAR-Free深度辅助学习模块的有效性. 虽然D$^4$LCN在一些指标上优于本文方法（仅采用LiDAR-Free深度图），但是冗余的模块设计带了一定程度的推理开销，本文方法与之相比在单张图像推理时间上快了5倍，参数量只有D$^4$LCN的22.5%，证明了LiDAR-Free深度辅助学习模块的高效性.

**表4　基于深度估计的单目3D目标检测方法在KITTI3D验证集上的结果比较**

| IoU@0.7 | 3D | | | BEV AP | | | 推理时间 | 参数量 |
|---|---|---|---|---|---|---|---|---|
| | Easy | Mod | Hard | Easy | Mod | Hard | (ms) | (M) |
| MonoGRNet*(AAAI19)[25] | 13.88 | 10.19 | 7.62 | 24.97 | 19.44 | 16.30 | 60 | 27.87 |
| MonoPSR†(CVPR19)[26] | 13.94 | 12.24 | 10.77 | 21.52 | 18.90 | 14.94 | 200 | 30.21 |
| D$^4$LCN*(CVPR20)[27] | **22.32** | 16.20 | 12.30 | **31.53** | **22.58** | 17.87 | 200 | 92.23 |
| MonoDLE†(CVPR21) | 16.93 | 14.07 | 11.99 | 23.58 | 19.34 | 17.49 | 40 | 21.49 |
| MonoDLE＋LiDAR-Free深度图† | 20.55 | **16.87** | **14.39** | 28.24 | 22.47 | **20.32** | **40** | **20.76** |

注：KITTI3D的验证集有两种划分方式split1、split2，本表中根据原论文标注了各结果采用的划分方式，本文实验采用split2

*：KITTI validation split1　†：KITTI validation split2

事实上，单目3D目标检测这一任务之所以难，一个重要原因就是难以基于单目图像进行准确的深度估计，而如果能够得到更为准确的深度，肯定能够提升3D目标检测的性能. 但是，单目网络来推理得到深度图本身就会带来一定的计算开销[28-29]，而本文Lidar-Free深度估计辅助学习模块推算出的深度信息是基于标注框的，计算简单，并不涉及模型的推理，只通过简单计算就可得到绝对尺度下的深度图.

### 4.4　与其他方法比较

在KITTI官方测试集上将本文提出的MonoAux与过去几年的优秀算法进行了比较. 其中，MonoDLE是本文基线方法，同时，MonoDLE、MonoFlex、GUPNet、MonoCon、GrooMeD-NMS、MonoRCNN、MonoRCNN＋＋、MonoDETR、MonoCD是几种标准单目3D检测器，这些检测器仅使用KITTI3D标注而不引入额外数据（CAD、LiDAR、外部私有数据等），这与本文方法类似，其中不乏在深度估计上提出新思路的方法（如MonoRCNN、MonoRCNN＋＋），也有利用几何特征的方法（如MonoDETR、MonoCD等）. AutoShape使用CAD模型. 除此之外，其余检测器均引入LiDAR数据，以引导单目3D检测器训练.

表5的结果显示，相比基线方法，本文性能提升显著，在3D AP的Mod指标上提升了4.64%（12.26% vs. 16.90）. 与其他方法相比，本文方法也具有显著的性能优势，明显优于一些不使用额外数据的方法，如MonoDLE、MonoFlex、GUPNet、MonoCon，甚至也优于部分使用额外数据的方法，如CaDDN、AutoShape、MonoDTR、MonoDistil等. 即使与最近顶级会议提出的方法，如MonoCD（CVPR24）相比也具有优势，在这些比较方法中，6个检测精度指标中本文方法有3个指标第1，且在3D目标检测最关注的Mod类目标上检测精度最优. 推理时间和参数量也满足实时处理的要求，具有较大优势.

### 4.5　3D目标检测结果可视化比较

为直观展示本文算法的有效性，图9(a1)、(b1)、(c1)、(d1)分别展示了基线模型MonoDLE、MonoDistill、DID-M3D与本文提出的MonoAux在

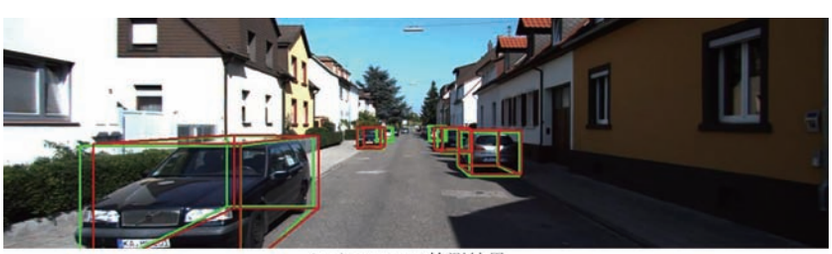
**表5 MonoAux与其他优秀方法在KITTI3D测试集上的结果**

| IoU@0.7 | 3D AP (Car *test*) | | | BEV AP (Car *test*) | | | 推理时间(ms) | 参数量(M) | 是否使用额外数据 |
|---|---|---|---|---|---|---|---|---|---|
| | Easy | Mod | Hard | Easy | Mod | Hard | | | |
| MonoDLE(CVPR21)[4] | 17.23 | 12.26 | 10.29 | 24.79 | 18.89 | 16.00 | 40 | 20.32 | 否 |
| MonoFLex(CVPR21)[14] | 19.94 | 12.89 | 12.07 | 28.23 | 19.75 | 16.89 | 30 | 20.84 | 否 |
| GrooMeD-NMS(CVPR21)[30] | 18.36 | 12.65 | 10.03 | 25.48 | 18.11 | 14.10 | 120 | 12.09 | 否 |
| GUPNet(ICCV21)[12] | 22.26 | 15.02 | 13.12 | 30.29 | 21.19 | 18.20 | 34 | 20.89 | 否 |
| MonoRCNN(ICCV21)[11] | 18.36 | 12.65 | 10.03 | 25.48 | 18.11 | 14.10 | 70 | - | 否 |
| MonoCon(AAAI22)[8] | 22.50 | 16.46 | 13.95 | 31.12 | 22.10 | 19.00 | 26 | 19.64 | 否 |
| MonoRCNN++(WACV23)[15] | 20.08 | 13.72 | 11.34 | - | - | - | - | 67.67 | 否 |
| MonoDETR(ICCV23)[16] | 25.00 | 16.37 | 13.58 | 33.60 | 22.11 | 18.60 | 38 | 21.47 | 否 |
| MonoCD(CVPR24)[10] | 25.53 | 16.59 | 14.53 | 33.41 | 22.81 | 19.57 | 36 | 54.25 | 否 |
| CaDDN(CVPR21)[31] | 19.17 | 13.41 | 11.46 | 27.94 | 18.91 | 17.19 | 630 | 20.82 | 是,LiDAR |
| AutoShape(ICCV21)[9] | 22.47 | 14.17 | 11.36 | 30.66 | 20.08 | 15.95 | 40 | 21.35 | 是,CAD |
| MonoDTR(CVPR22)[18] | 21.99 | 15.39 | 12.73 | 28.59 | 20.38 | 17.14 | 37 | 69.79 | 是,LiDAR |
| MonoDistill(ICLR22)[32] | 22.97 | 16.03 | 13.60 | 31.87 | 22.59 | 19.72 | 40 | 37.68 | 是,LiDAR |
| DID-M3D(ECCV22)[13] | 24.40 | 16.29 | 13.75 | 32.95 | 22.76 | 19.83 | 40 | 21.76 | 是,LiDAR |
| MonoAux(本文方法) | 23.87 | 16.90 | 14.09 | 32.30 | 23.00 | 19.84 | 40 | 20.76 | 否 |

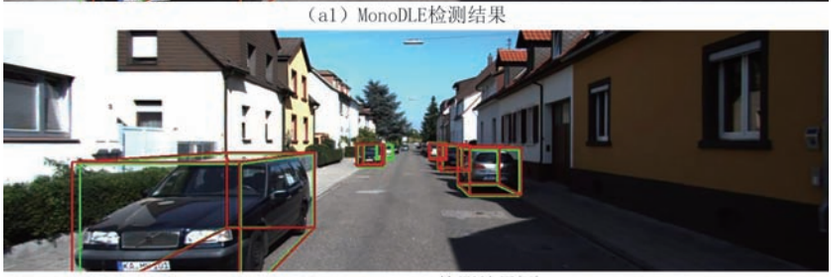
KITTI3D验证集上的3D目标检测结果对比图.其中,绿色3D检测框表示KITTI数据集标注结果,红色检测框则表示检测器的预测结果.图9(a2)-(d2)分别展示了对应的三维可视图,从BEV视角去看远处目标距离为49.29 m.得益于辅助学习模块,本文方法绿色框标注结果与红色框预测结果更为贴合,无误检和漏检,MonoAux实现了更加准确的深度估计以及目标分类,尤其是对较远距离目标的成功检测.

### 4.6 Loss函数收敛曲线比较

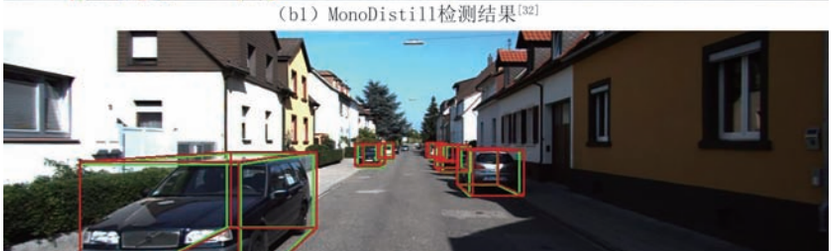
需要说明的是,本文的核心思路是利用变换得到的额外信息作为辅助,以进行监督训练提升检测性能.无论是投影点辅助训练分支中利用3D检测框计算10个角点投影点以提升3D检测框预估的可



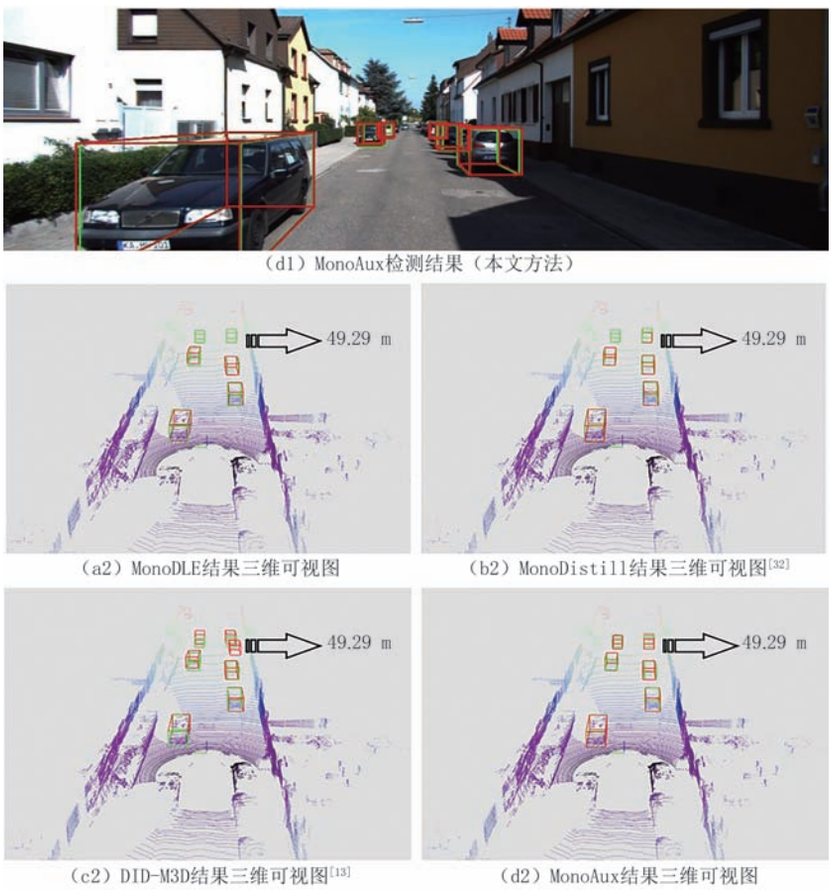
(a1) MonoDLE检测结果



(b1) MonoDistill检测结果[32]



(c1) DID-M3D检测结果[13]




（d1）MonoAux检测结果（本文方法）

（a2）MonoDLE结果三维可视图　（b2）MonoDistill结果三维可视图[32]

（c2）DID-M3D结果三维可视图[13]　（d2）MonoAux结果三维可视图

图9　MonoAux与MonoDLE、MonoDistill、DID-M3D检测结果可视化对比图

靠性，还是增加对3D框中心投影点与角点投影点之间的偏移量预测，间接为角点投影点预测、其他关键点偏移量预测提供额外的监督信号，亦或是将单个关键点深度替换为目标各个区域较为准确的深度估计结果，并将其作为监督信号辅助训练，均是基于这一思路，而其有效性也可以从图10中得到验证，从不同迭代次数下检测分支Loss函数曲线的变化可以看出，本文MonoAux在MonoDLE的基础上增加了投影点辅助学习、LiDAR-Free深度估计辅助学习，损失函数收敛更快，辅助信息起到了有效的监督训练的作用.

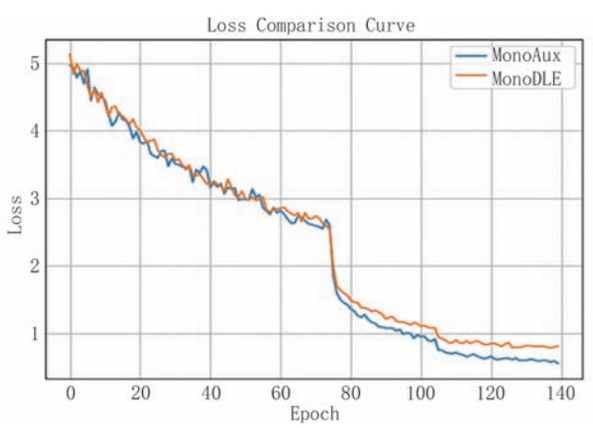



图10　MonoAux与MonoDLE检测分支Loss函数收敛曲线

## 5　总　结

本文提出了一种多关键点约束与深度估计辅助的单目3D目标检测算法MonoAux，解决现有基于关键点检测范式的单目3D目标检测器存在的监督信号不足的问题. 设计了两种新颖的辅助学习模块，其中，投影点辅助学习模块通过引入更多关键点约束来预测目标，使得检测器的关键点预测分支得到更好的学习，提高投影点预测的准确性；而LiDAR-Free深度估计辅助学习模块通过本文首次提出的LiDAR-Free解耦深度估计算法，在无需额外借助激光雷达的情况下生成目标区域的稠密深度，从而为单目3D目标检测器提供更多深度估计的监督信号，进一步提升深度估计精度. 这两种模块仅在训练阶段使用，推理阶段不引入额外的计算成本，因此保证了推理阶段的检测效率.

实验结果验证了本文方法的有效性. 在

KITTI3D目标检测验证集和测试集上，相较于MonoDLE基线网络，提出的MonoAux算法在目标检测精度上分别提升了3.87%和4.64%. 与近年顶级会议提出的MonoFLex、GUPNet、MonoCon、CaDDN、AutoShape、MonoDTR、MonoDistill、DID-M3D等方法相比，本文也具有显著的性能优势，可在不引入额外激光点云数据的前提下取得与激光雷达相近、甚至更优的性能表现.

## 参 考 文 献


[1] Li P, Jin J. Time3D: End-to-end joint monocular 3D object detection and tracking for autonomous driving//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 3885-3894

[2] Jiang Jun-Jun, Li Zhen-Yu, Liu Xian-Ming. Deep learning based monocular depth estimation: A survey. Chinese Journal of Computers, 2022, 45(6): 1276-1307 (in Chinese)
(江俊君，李震宇，刘贤明. 基于深度学习的单目深度估计方法综述. 计算机学报，2022，45(6)：1276-1307)

[3] Zhou X, Wang D, Krähenbühl P. Objects as points. arXiv preprint arXiv:1904.07850, 2019

[4] Ma X, Zhang Y, Xu D, et al. Delving into localization errors for monocular 3D object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 4721-4730

[5] Yu F, Wang D, Shelhamer E, et al. Deep layer aggregation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2403-2412

[6] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation//Proceedings of the 14th European Conference of Computer Vision-ECCV 2016. Amsterdam, The Netherlands, 2016: 483-499

[7] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2117-2125

[8] Liu X, Xue N, Wu T. Learning auxiliary monocular contexts helps monocular 3D object detection//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2022: 1810-1818

[9] Liu Z, Zhou D, Lu F, et al. Autoshape: Real-time shape-aware monocular 3D object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 15641-15650

[10] Yan L, Yan P, Xiong S, et al. MonoCD: Monocular 3D object detection with complementary depths//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 10248-10257

[11] Shi X, Ye Q, Chen X, et al. Geometry-based distance decomposition for monocular 3D object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 15172-15181

[12] Lu Y, Ma X, Yang L, et al. Geometry uncertainty projection network for monocular 3D object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 3111-3121

[13] Peng L, Wu X, Yang Z, et al. DID-M3D: Decoupling instance depth for monocular 3D object detection//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 71-88

[14] Zhang Y, Lu J, Zhou J. Objects are different: Flexible monocular 3D object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Montreal, Canada, 2021: 3289-3298

[15] Shi X, Chen Z, Kim T K. Multivariate probabilistic monocular 3D object detection//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Hawaii, USA, 2023: 4281-4290

[16] Zhang R, Qiu H, Wang T, et al. MonoDETR: Depth-guided transformer for monocular 3D object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 9155-9166

[17] Huang K C, Wu T H, Su H T, et al. Monodtr: Monocular 3D object detection with depth-aware transformer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 4012-4021

[18] Li Y, Ge Z, Yu G, et al. Bevdepth: Acquisition of reliable depth for multi-view 3D object detection//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2023: 1477-1485

[19] Xu C, Wu B, Hou J, et al. Nerf-det: Learning geometry-aware volumetric representation for multi-view 3D object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Vancouver, Canada, 2023: 23320-23330

[20] Zhou Y, Tuzel O. Voxelnet: End-to-end learning for point cloud based 3D object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4490-4499

[21] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014

[22] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[23] Ku J, Harakeh A, Waslander S L. In defense of classical image processing: Fast depth completion on the CPU// Proceedings of the 15th Conference on Computer and Robot Vision. Toronto, Canada, 2018: 16-22

[24] Hu M, Wang S, Li B, et al. Penet: Towards precise and efficient image guided depth completion// Proceedings of the 2021 IEEE International Conference on Robotics and Automation. Xi′an, China, 2021: 13656-13662

[25] Qin Z, Wang J, Lu Y. Monogrnet: A geometric reasoning

network for monocular 3D object localization//Proceedings of the 2019 AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 8851-8858

[26] Ku J, Pon A D, Waslander SL. Monocular 3D object detection leveraging accurate proposals and shape reconstruction// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 11867-11876

[27] Ding M, Huo Y, Yi H, et al. Learning Depth-guided convolutions for monocular 3D object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA, 2020: 1000-1001

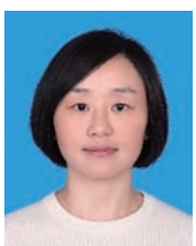
[28] Ranftl R, Lasinger K, Hafner D, et al. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(3): 1623-1637

[29] Yang L, Kang B, Huang Z, et al. Depth anything: Unleashing the power of large-scale unlabeled data. arXiv preprint arXiv: 2401.10891, 2024

[30] Kumar A, Brazil G, Liu X. Groomed-nms: Grouped mathematically differentiable nms for monocular 3D object detection//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 8973-8983

[31] Reading C, Harakeh A, Chae J, et al. Categorical depth distribution network for monocular 3D object detection// Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 8555-8564

[32] Chong Z, Ma X, Zhang H, et al. Monodistill: Learning spatial features for monocular 3D object detection. arXiv preprint arXiv: 2201.10830, 2022



**ZHENG Jin**, Ph. D., associate professor. Her main research interests are computer vision, object detection and tracking.

**WANG Sen**, M. S. His main research interests are object detection and computer vision.

**LI Hang**, Ph. D. candidate. His main research interests are object detection and computer vision.

**ZHOU Yu-Hai**, M. S. candidate. His main research interests are computer vision, image & video processing.

**Background**

3D object detection can provide information such as object center point coordinates, object size and deviation angle, making it an indispensable capability for autonomous driving, robot navigation, etc. Compared to binocular cameras, monocular cameras have advantages such as simple structure, low economic cost, and convenient calibration, therefore, monocular 3D object detection has received increasing attention. However, due to the lack of accurate depth estimation, the accuracy of 3D object detection methods without LiDAR point clouds is still very low, which cannot meet the needs of practical applications.

The existing mainstream monocular 3D object detection algorithms adopt the keypoint-based paradigm, and this paradigm results in inaccurate key-point prediction and depth estimation, which hinder the performance of monocular 3D detectors. In order to overcome the bottleneck issues of inaccurate keypoint-based prediction and depth estimation, which affect the accuracy of monocular 3D object detection, this paper proposes a novel monocular 3D detector called MonoAux with multi-keypoint constraints and depth estimation assistance. MonoAux adopts the corner projection points of the 3D detection bounding box, as well as the center projection points of the upper and lower surfaces as supplements for the center projection points of the 3D bounding box, and thus the constraint of multiple keypoints improves the precision of keypoint prediction. Additionally, a LiDAR-Free decoupling depth estimation method is proposed to enhance the accuracy of depth estimation, which provides more auxiliary supervision signals for depth estimation by exploiting geometric relationships, even without introducing additional LiDAR point cloud data. Mult-keypoint constraints and depth estimation assistance are only used during the training phase, and do not introduce additional computational costs during the inference phase. The results on the KITTI3D object detection validation set and test set show that the proposed MonoAux algorithm improves 3D object detection accuracy by 3.87% and 4.64% compared to the baseline network MonoDLE. Furthermore, compared with other SOTA methods, the proposed method also has significant performance advantages, even better than some methods that use additional data.

Our group is investigating 3D object detection. Aiming at the low accuracy of existing binocular stereo matching and depth estimation methods, we propose a multi-scale binocular stereo matching network based on semantic association. In view of the

accuracy of existing 3D object detection algorithms based on Pseudo-LiDAR is far lower than that based on real LiDAR, we study the reconstruction of Pseudo-LiDAR and proposes a 3D object detection algorithm suitable for Pseudo-LiDAR. Furthermore, we propose MonoSKD, a novel knowledge distillation framework for monocular 3D detection based on Spearman correlation coefficient, to learn the relative correlation between cross-modal features. Our method achieves state-of-the-art performance until submission with no additional inference computational cost. In fact, the combination of MonoAux proposed in this paper and MonoSKD achieved Rank 1 on the official KITTI3D test set when it was submitted.