DOI: 10.13382/j.jemi.B2306361

# 联合多注意力和 C-ASPP 的单目 3D 目标检测*

郑自立 徐 健 刘秀平 刘高峰 赵一剑 夏代洪
(西安工程大学电子信息学院 西安 710048)

**摘 要**:针对单目 3D 检测中网络结构复杂、深度估计后得到的目标深度信息不精确的问题,本文提出一种端到端的联合多注意力深度估计的单目 3D 目标检测网络结构(CDCN-3D)。首先,为获取目标显著特征,引入自适应空间注意力机制,对像素特征进行聚集,以增强局部特征来提升网络表征能力;其次,为改善深度估计时局部信息丢失问题,利用改进 C-ASPP 使每个深度信息都能够捕获更加精确的方向感知和位置敏感信息;最后,利用精确的 P-BEV 将得到的目标三维信息映射到二维平面,再用单级目标检测器完成检测输出任务。实验结果证明,CDCN-3D 网络在 KITTI 数据集上,在 FPS 与现有单目 3D 检测网络持平情况下,其准确率优于其他网络,在 Car、Pedestrian、Cyclist 类中,其检测精确度分别提升 2.31%、1.48%、1.14%,能够完成 3D 目标检测任务。

**关键词**: 单目 3D 目标检测;深度估计;多注意力机制;机器视觉;自动驾驶
**中图分类号**: TN966.6 **文献标识码**: A **国家标准学科分类代码**: 540.10

# Combined multi-attention and C-ASPP network for monocular 3D object detection

Zheng Zili Xu Jian Liu Xiuping Liu Gaofeng Zhao Yijian Xia Daihong
(School of Electronics and Information, Xi'an Polytechnic University, Xi'an 710048, China)

**Abstract**: In monocular 3D detection, the complex network structure and inaccurate target depth information obtained after depth estimation are two problems that need to be dealt with. To address this issue, we propose an end-to-end joint multi-attention depth estimation monocular 3D target detection network structure, named CDCN-3D. First of all, to obtain the salient features of the target, we introduce an adaptive spatial attention mechanism to aggregate the pixel features, which enhances local features and improves the network representation ability. Second, we use an improved C-ASPP approach to address the problem of local information loss in depth estimation, capturing more accurate direction perception and position-sensitive information for each depth information. Finally, the accurate P-BEV is used to map the three-dimensional information of the target to a two-dimensional plane, and then the single-stage target detector is used to complete the detection and output task. Through experiments on the KITTI dataset, the proposed CDCN-3D network shows improved accuracy compared to other networks, with the same FPS as that of the existing monocular 3D detection network. More specifically, and the detection accuracy of the CDCN-3D network is improved by 2.31%, 1.48%, 1.14% respectively by the class of Car、Pedestrian、Cyclist, which can complete the 3D target detection task.

**Keywords**: monocular 3D target detection; depth estimation; multi-attention mechanism; machine vision; autonomous driving

## 0 引 言

目前,在二维目标检测领域,最先进的检测算法识别能力已远超人眼[1]。但二维层面的检测算法只能得到物体的类别和位置信息,无法感知目标深度信息;在实际应用中,特别是自动驾驶领域,需要对物体在其真实场景中的 3D 信息进行提取,二维目标检测的效果与工作任务要求相差甚远[2]。

大多数 3D 检测依赖于激光雷达系统[3]或立体相机[4]来提供场景的深度信息作为输入完成检测任务,能够准确地完成检测任务,但检测系统复杂增加其实现成

收稿日期: 2023-03-23 Received Date: 2023-03-23
*基金项目:陕西省科技厅项目(2018GY-173)、西安市科技局项目(GXYD7.5)资助

本。近年来单目 3D 目标检测方法因其硬件成本低、功耗小以及易部署于现实场景中等优点而受到广泛关注[5]，单目图像二维目标检测算法相对成熟，可以快速实现目标在图像平面的定位和分类[6]，在三维目标检测中仍然存在局限性，尤其是针对室外场景对目标的深度估计更加困难，通过单一的纹理特征[7]难以确定目标物体在三维目标空间信息。常见的单目三维检测算法需要结合几何特征[8]、三维模型匹配[9]、先验信息融合[10]、深度估计网络[11]等方法回归目标的三维几何特征。随着自动驾驶和增强现实等技术的发展，基于单目视觉的三维目标检测算法已经成为研究的重点[12]。

现有单目 3D 目标检测方法主要有 3 种，分别为直接回归法、深度图法和网格法。直接法是从图像中直接估计出 3D 检测框，无需预测中间的 3D 场景表示，结合 2D 图像平面和 3D 空间的几何关系来辅助检测。但 3D 检测框(3D box)从 2D 图像中生成时没有得到检测物体明确的深度信息，相对于其他方法定位性能较差[13]；基于深度的方法先利用深度估计网络估计出图像的像素级深度图，将该深度图作为输入用于 3D 目标检测任务，因其深度估计子网络和目标检测子网络分离训练的结构，导致隐含信息丢失检测误差增大[14]；基于网格的方法通过预测出鸟瞰图(birds eye vieus，BEV)网格替代深度估计作为 3D 检测输入，通过采样图像特征来填充体素网格然后将体素投射到图像平面，并将其转换为 BEV 表示。这种方法由于多个体素可以投影到同一图像特征上，导致特征沿着投影射线重复出现，检测精度降低[15]。

单目检测方法无法直接获取场景的深度信息且将场景信息投影到二维图像平面时，会丢失部分场景深度信息[16]，因此其检测基准性能明显落后于激光雷达和立体方法[17]。为解决这类问题，通常在训练阶段采用深度估计来学习场景深度[18]，利用一张或者唯一视角下的二维图像来估计出图像中每个像素距拍摄源的距离[19]。基于深度学习的单目估计是依据像素值关系来反映深度关系，方法是拟合一个函数把图像映射成深度图，从单张图片中获取具体的深度相当于从二维图像推测出三维空间。Saxena 等[20]在最大化后验概率框架下以超像素为单元，利用马尔可夫随机场拟合特征与深度、不同尺度的深度之间的关系，实现对深度的估计；Liu 等[21]将深度卷积神经网络与连续条件随机场结合，提出深度卷积神经场从单幅图像中估计深度；Wang 等[22]提出了一个统一的框架联合深度估计和语义分割任务，将原图分割成各个区域，在全局预测的指导下得到区域级的深度和语义标签，利用两层的条件随机场由像素级的全局预测和区域级的局部预测得到最后的优化结果。如何对图像进行精确的深度估计成为单目 3D 物体检测的难点[23]；此外，在对图像中的深度信息进行折叠时，一般采取先将得到带有深度信息图像特征转换到 3D 空间并最终转换到 BEV[24]网格来进行最终的检测任务[25]，但在 BEV 学习过程中图像中相似的图像特征可能会被投影于多个位置，造成特征模糊[26]，增加在场景中对物体进行定位难度。

针对上述单目 3D 物体检测中存在的问题，本文提出一种用于单目 3D 物体检测的端到端联合多注意力和深度估计的网络模型(CDCN-3D)。为能够更好获取图像特征，利用自适应空间注意力机制(SAM)来对目标中的特征进行聚集，增加特征之间的判别性；通过改进空洞金字塔结构(C-ASPP)解决局部信息丢失问题，从而提取编码器特征信息生成高质量深度图；引入透视鸟瞰图(P-BEV)投影来保留物体的大小和深度信息，将物体的空间信息映射到平面世界，使利用单级检测器完成最后的检测任务。CDCN-3D 以端到端学习的方式联合深度估计和三维目标检测，有效地提高目标检测效果。

## 1 基本原理

该网络结构以 RGB 图像作为输入，利用特征提取网络获取图像二维特征后与深度估计信息进行融合，生成截锥特征网络并投影到体素空间[27]，然后将其折叠生成 P-BEV 图。最后，利用高效的单级目标检测器来完成检测输出任务，实现 3D 对象检测，网络整体结构如图 1 所示。

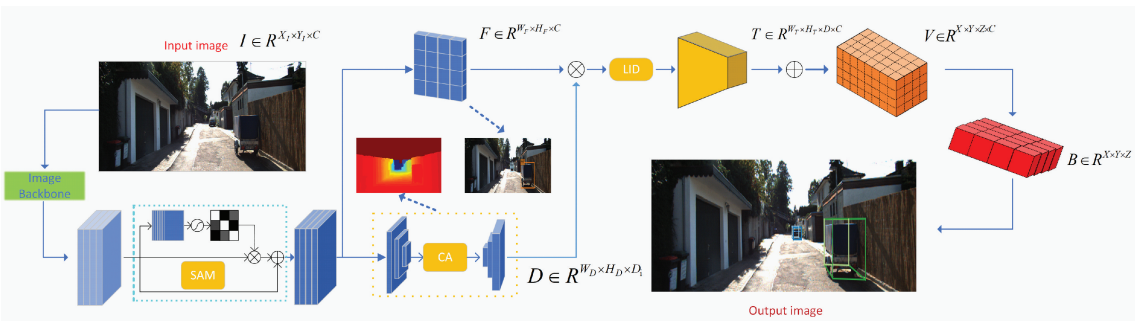


图 1 CDCN-3D 网络结构

Fig. 1 CDCN-3D network structure diagram

图 1 中，$\boldsymbol{I}$ 是输入的图像大小，$\boldsymbol{X}_I \times \boldsymbol{Y}_I$ 是输入图像大小，$C$ 为特征通道数；$\boldsymbol{F}$ 为提取的图像特征，$\boldsymbol{W}_F \times \boldsymbol{H}_F$ 是二维图像特征大小；$D$ 是深度估计所得目标深度信息，$\boldsymbol{W}_D \times \boldsymbol{H}_D$ 是二维图像特征大小，$D_i$ 为深度索引；$\boldsymbol{T}$ 为多维信息融合后的截锥特征，$\boldsymbol{W}_T \times \boldsymbol{H}_T \times \boldsymbol{D}$ 为特征信息位置；$\boldsymbol{V}$ 是映射所得的体素网络，$\boldsymbol{X} \times \boldsymbol{Y} \times \boldsymbol{Z}$ 为特征坐标；$\boldsymbol{B}$ 为经过 P-BEV 处理得到的检测网络输入，$\boldsymbol{X} \times \boldsymbol{Y} \times \boldsymbol{Z}$ 为特征信息坐标。

### 1.1　优化三维特征获取

对于输入的 RGB 图像，利用特征提取网络和深度估计网络处理融合获得图像三维特征。首先，输入 RGB 图片经过一个 ResNet50 的 backbone 进行特征提取，得到 $\boldsymbol{W} \times \boldsymbol{H} \times \boldsymbol{C}$ 大小的特征图，其中，$\boldsymbol{W} \times \boldsymbol{H}$ 为图像特征表示的高和宽，$C$ 为特征通道数，然后经过两个并行的分支，其中一个分支将特征通道 $C$ 由 256 降至 64；另一个分支用于估计特征深度信息，深度估计网络采用 DeepLabV3+得到 $\boldsymbol{W} \times \boldsymbol{H} \times \boldsymbol{D}$ 的深度估计结果，其中 $D$ 是预测深度。

1）改进特征提取网络

图片输入后，使用 ResNet-50 主干网络进行特征提取。ResNet-50 网络使用残差结构，允许特征提取网络加深，避免“网络退化”现象，残差结构如图 2 所示。

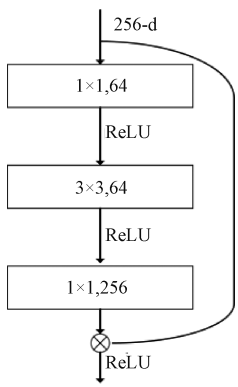


图 2　残差网络结构

Fig. 2　Residual network structure

为得到空间高分辨率的特征图，协助网络模型在复杂背景下更好地区分所需对象目标，在网络特征提取的过程中引入 SAM。由于特征图上的像素点之间存在一定的相关性，当像素点的距离增大时，它们之间的相关性将会降低甚至失去联系。因此，通过 SAM 模块对目标中的像素进行聚集，增强局部特征，提升特征表征能力，从而更好地预测出不同深度的目标对象。

空间注意力模块的运算公式过程如式(1)所示：

$$M_s(F) = \delta(f^{5\times5}[AvgPool(F); MaxPool(F)]) \quad (1)$$

式中：$\boldsymbol{F}$ 是经过通道注意力机制得到的特征图，$\delta$ 为 sigmoid 激活函数，$f^{5\times5}$ 表示卷积核为 5×5 大小的卷积层，$AvgPool(F)$，$MaxPool(F)$ 表示经过池化操作后的特征图。其结构如图 3 所示。

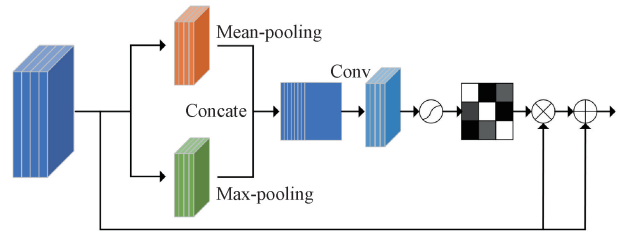


图 3　SAM 网络结构

Fig. 3　SAM network structure

2）改进 C-ASPP

特征提取后，对图像特征进行逐像素深度估计，得到其深度分布 $D \in R^{W_D \times H_D \times D_i}$，其中 $D_i$ 为通过深度估计深度信息。使用语义分割 DeepLabV3+来预测图像特征的深度信息，使图像特征的每个像素上预测其落入深度离散化后的深度区间的概率。

DeepLabV3+中使用 ASPP 来完成准确、高效地对任意尺度的区域进行分类。ASPP 中使用空洞卷积(atrous convolution)，是一种增加感受野的方法，空洞卷积是为解决语义分割中，输出图像的 size 要求和输入图像的 size 一致而需要上采样，但使用池化操作来增大感受野会降低分辨率，导致上采样无法还原由池化导致的一些细节信息的损失的问题而提出的。为了减小这种损失，需要移除池化层，因此空洞卷积应运而生。ASPP 模型如图 4 所示

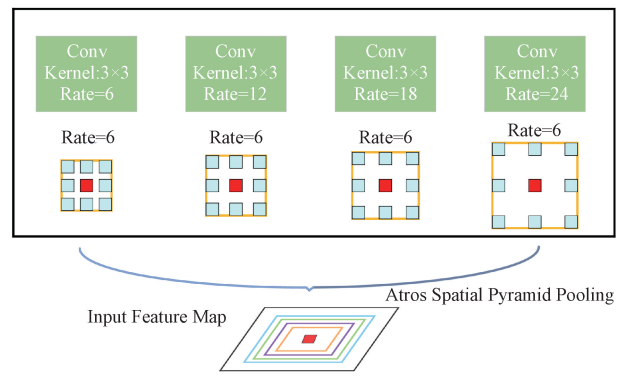


图 4　ASPP 模型

Fig. 4　ASPP model

由于空洞卷积的计算方式类似于棋盘格式，某一层得到的卷积结果，来自上一层的独立的集合，没有相互依赖，因此该层的卷积结果之间没有相关性，即局部信息存在丢失，为改善 ASPP 造成的局部信息丢失的问题，利用改进 C-ASPP 以获得多尺度信息，使每个深度信息都能够捕获更加精确的方向感知和位置敏感信息，具体来说，

对输入分别进行 $X$ 和 $Y$ 方向池化后进行 concat 后利用 1×1 卷积核进行降维和激活，再沿着空间维度对其进行 *split* 操作后使用 *sigmoid* 激活函数进行激活，其过程如图 5 所示。

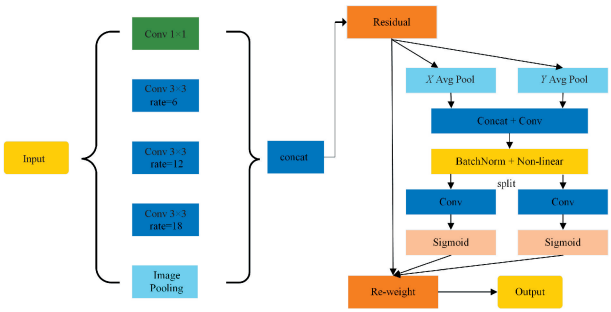


图 5　改进 C-ASPP 结构

Fig. 5　Improved C-ASPP structure

同时，对图像特征执行通道缩减操作，用来生成最后融合所需要的图像特征。即通过 ResNet-50 主干网络提取得到特征通道数为 256，使用 1×1 的卷积核进行下采样操作，将通道数降为 64，可以在保持特征图尺度不变的前提下大幅增加非线性特性，使得特征图表达能力更佳。

3）多维信息融合

得到深度估计图 $\boldsymbol{D}$ 和通道缩减图像特征 $\boldsymbol{F}$ 后，通过作外积生成截锥网络。设 $(m,n,c)$ 表示图像特征 $\boldsymbol{F}$ 的坐标，$(m,n,d_i)$ 表示深度估计中 $\boldsymbol{D}$ 的坐标，其中 $(m,n)$ 为像素坐标，$c$ 为特征通道，$di$ 为预测深度概率所属空间的索引，如式(2)所示：

$$T(m,n)=F(m,n)\otimes D(m,n) \tag{2}$$

式中：$\boldsymbol{D}(\boldsymbol{m},\boldsymbol{n})$ 为估计所得的深度分布，$\boldsymbol{T}(\boldsymbol{m},\boldsymbol{n})$ 为每个像素外积操作后对应生成的截锥特征，如图 6 所示。

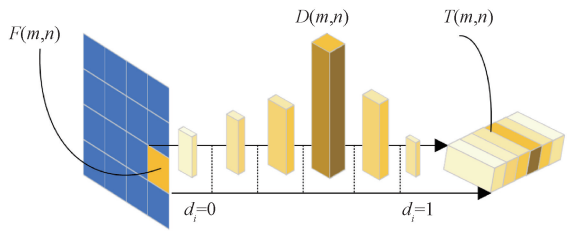


图 6　生成截锥特征

Fig. 6　Generating truncated cone features

## 1.2　P-BEV 特征生成

BEV 网格是从 RGB 图像中预测 3D 场景的一种可视化地图，在获得 BEV 视图前，需要将空间分割成体素，利用体素对点云进行下采样，将每个体素作为一个点进行投影。

为更好的完成 3D 目标检测任务，在网络中引入 P-BEV，其能在二维平面上保留物体的大小和深度信息，同时，其卷积的过程能够有效地解决特征采样造成的信息损失和结构失真，更有效的完成检测任务。

其生成过程为将直接堆叠体素特征 $\boldsymbol{V}\in\boldsymbol{R}^{X\times Y\times Z\times C}$，就可以得到 P-BEV 特征 $\boldsymbol{B}\in\boldsymbol{R}^{X\times Y\times C}$。具体的操作过程是：将 $Z$ 轴和 $C$ 轴拼接起来，然后采用卷积池化操作将 $Z\times C$ 通道降维至 $C$，就得到了 P-BEV 特征 $\boldsymbol{B}$，其过程如图 7 所示。P-BEV 网格在极大的提高了计算性能的同时，具备跟体素网格法相似的检测性能。

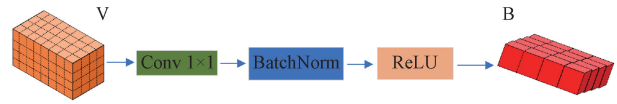


图 7　P-BEV 生成

Fig. 7　P-BEV generation

## 1.3　损失函数

考虑到 3D 目标的复杂度及周围环境对检测性能的影响，在网络中利用不同的损失函数来使网络模型提高识别精度。

在深度估计中，需要对图像的深度信息进行分布预测并得到其分布概率，同时使用离散化方法将尺寸为 $\boldsymbol{W}_F\times\boldsymbol{W}_H$ 的深度图转化为面元索引，随后使用独热编码生成深度分布标签，使用焦点损失函数，如式(3)所示：

$$L_{\text{depth}}=\frac{1}{W_F\cdot W_H}\sum^{W_F}\sum^{W_H}FL(D,\hat{D}) \tag{3}$$

式中：$D$ 为深度分布预测，$\hat{D}$ 为深度分布标签。

为解决 P-BEV 中因目标对象与环境背景占比相差较大引起正负样本比重不平衡的问题，采用多任务分类损失函数 *Focal Loss* 作为网络的分类损失函数，如式(4)所示：

$$L_{\text{cls}}(p,y)=\begin{cases}-\alpha(1-p)^{\gamma}\log p, y=1\\-(1-\alpha)p^{\gamma}\log(1-p), \text{其他}\end{cases} \tag{4}$$

式中：$\alpha$ 和 $\gamma$ 为函数的可调参数，设置为 $\alpha=0.25$，$\gamma=2$；$p$ 和 $y$ 网络输出的分类置信度和对应真值。

而对于回归损失函数，3D B-Box 被建模为一个 7 维向量表示，分别为 $(x,y,z,w,h,l,\theta)$，其中 $(x,y,z)$ 为坐标中心，$w,h,l$ 为尺寸数据，$\theta$ 为方向角，7 个变量采用 Smooth L1 损失进行回归训练，如式(5)所示：

$$L_{\text{loc}}=\sum_{b\in(x,y,z,w,h,l,\theta)}\text{Smooth}L1(\Delta b) \tag{5}$$

式中：$\Delta b$ 为各个变量的偏移量，偏移量的计算如式(6)~(8)所示：

$$\Delta x=\frac{x^{gt}-x^{a}}{d^{a}},\Delta y=\frac{y^{gt}-y^{a}}{d^{a}},\Delta z=\frac{z^{gt}-z^{a}}{d^{a}} \tag{6}$$

$$\Delta w=\log\frac{w^{gt}}{w^{a}},\Delta l=\log\frac{l^{gt}}{l^{a}},\Delta h=\log\frac{h^{gt}}{h^{a}} \tag{7}$$

$$\Delta\theta = \sin(\theta^{gt} - \theta^{a}) \tag{8}$$

式中：$x^{gt}$ 和 $x^{a}$ 分别是 groud truth 和 anchor box，如式(9)所示：

$$d^{a} = \sqrt{(w^{a})^{2} + (l^{a})^{2}} \tag{9}$$

由于角度定位损失无法区分翻转框，因此在离散方向即 $L_{\text{dir}}$ 使用 softmax 作为损失函数。因此网络模型的总体损失函数如式(10)所示：

$$L = \lambda_{\text{depth}} L_{\text{depth}} + \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{loc}} L_{\text{loc}} + \lambda_{\text{dir}} L_{\text{dir}} \tag{10}$$

式中：$\lambda_{\text{depth}}$、$\lambda_{\text{cls}}$、$\lambda_{\text{loc}}$、$\lambda_{\text{dir}}$ 为损耗加权因子。

## 2　实验与分析

为了证明 CDCN-3D 网络的有效性，实验在 KITTI 3D 开放数据集上进行测试和验证。KITTI 3D 开放数据集分为 7 481 个训练样本和 7 518 个测试样本。本次实验也将训练样本分为训练集(3 712 个样本)和验证集(3 769 个样本)，通过在 train 和 val 集上训练深度网络模型，并将 CDCN-3D 与现有方法进行比较。

本文的实验平台环境主要以实验使用 Linux 操作系统，CPU 为 i7-7800X，显卡 GPU 为两张 NVIDIV GTX2080Ti，64 GB 内存。采用 Pycharm 作为开发工具编写算法软件。训练环境基于 PyTorch、Cuda10.2 和 Cudnn7.6.5 的深度学习框架，编程语言采用 Python，实验训练 100epochs。

### 2.1　KITTI 数据集对比实验结果

在 KITTI 数据集的结果用平均精度(AP)来进行评估网络模型，AP 是指回率从 0~1 间全部值对应精确度的平均值，其表达式如式(11)所示，而在实际应用中，通常将其描述在每个可能的阈值处的精确度与召回率的变化值积之和，如式(12)所示：

$$AP = \int_{0}^{1} p(r)\,\mathrm{d}r \tag{11}$$

$$AP = \sum_{k=1}^{N} p(k)\Delta r(k) \tag{12}$$

式中：$p(r)$ 为精确度随召回率变化曲线，$p(k)$ 为精确度值，$\Delta r(k)$ 为召回率的变化值。

KITTI 数据集分为汽车、行人和骑行者 3 类，并根据数据集中每种类别的识别简易程度分为简单(特征明显)、中等(特征遮挡)和困难(特征模糊)3 个级别分别进行测试。将汽车类别的 IoU 设置为 0.7，行人和骑行者类别的 IoU 设置为 0.5，结果如表 1 所示。可以看到，CDCN-3D 网络在汽车类和行人类的简单、中等和困难 3 个层次都有了显著的提升。在检测难度相对较低的 Car 类中，其平均精度的提升超过了 2.31%。而在检测难度较高的骑行者类别上，其检测精度提升 1.14%，能有效的完成检测任务。

**表 1　不同算法在 KITTI 数据集上测试结果**

**Table 1　Test results of different algorithms on KITTI dataset**　(%)

| 算法 | Car (IOU=0.7) | | | Pedestrian (IOU=0.5) | | | Cyclist (IOU=0.5) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Easy | Mod. | Hard | Easy | Mod. | Hard | Easy | Mod. | Hard |
| OFT | 1.61 | 1.32 | 1.00 | 0.63 | 0.36 | 0.35 | 0.14 | 0.06 | 0.07 |
| M3D-RPN | 14.76 | 9.71 | 7.42 | 4.92 | 3.48 | 2.94 | 0.94 | 0.65 | 0.47 |
| MonoPair | 13.04 | 9.99 | 8.65 | 10.02 | 6.68 | 5.53 | 3.79 | 2.12 | 1.83 |
| MoVi-3D | 15.19 | 10.90 | 9.26 | 8.99 | 5.44 | 4.57 | 1.08 | 0.63 | 0.70 |
| D4LCN | 16.65 | 11.72 | 9.51 | 4.55 | 3.42 | 2.83 | 2.45 | 1.67 | 1.36 |
| CaDDN | 19.17 | 13.41 | 11.46 | 12.87 | 8.14 | 6.76 | 7.00 | 3.41 | 3.30 |
| DFR-NET | 19.40 | 13.63 | 10.35 | 6.09 | 3.62 | 3.39 | 5.69 | 3.58 | 3.10 |
| DDMP-3D | 19.71 | 12.78 | 9.8 | 4.93 | 3.55 | 3.01 | 4.18 | 2.5 | 2.32 |
| MonoFlex | 19.94 | 13.89 | 12.07 | 9.43 | 6.31 | 5.26 | 4.17 | 2.35 | 2.04 |
| MonoEF | 21.29 | 13.87 | 11.71 | 4.27 | 2.79 | 2.21 | 1.8 | 0.92 | 0.71 |
| GUPNet | 22.26 | 15.02 | 13.12 | 14.95 | 9.76 | 8.41 | 5.58 | 3.21 | 2.66 |
| CDCN-3D(本文) | 24.57 | 16.78 | 14.41 | 16.43 | 10.68 | 8.26 | 8.14 | 3.71 | 2.97 |
| Improvements | +2.31 | +1.76 | +1.29 | +1.48 | +0.92 | −0.15 | +1.14 | +0.50 | −0.71 |

### 2.2　消融实验

1)网络消融对比使用

CDCN-3D 网络使用多注意力机制来获得更有效的深度信息，并将深度特征投影到 BEV 空间中，来获得更高准确的检测精度。为验证各个模块的有效性，本文设计了消融实验并根据 *AP* 指标数据展示其对网络的性能提升。

为验证 SAM 对网络特征提取能力的优化，将优化网络与 Basebone 网络进行对比实验，其结果如表 2 中实验 1、实验 2 所示。数据证明，在利用 SAM 对特征对目标中的像素特征进行聚集而增强局部特征，能够有效地提升网络检测精度，其在 Car(IoU=0.7)实验中，Easy 、Mod 以及 Hard 均有提升，特别是在 Easy，其精度提升 0.45%；在验证 C-ASPP 对获取深度信息的提升效果，将 Basebone 中 ASPP 与 C-ASPP 进行对比实验，其结果如表 2 中实验 2、实验 3 所示。数据证明，C-ASPP 能够让每个深度信息

都能够捕获更加精确的方向感知和位置敏感信息,使网络更好的获得特征深度信息,实现更加精确的目标检测任务;为验证两种深度离散方式 SID 和 LID 对网络检测能力提升的不同效果,其结果如表 2 实验 6 以及最终实验所示,实验数据表明,SID 深度离散方式在一定程度上能够对深度数值进行离散化从而优化网络从而提升检测精度,但是相比于 LID 离散化方式,因其不能够为所有的深度信息提供平衡的深度估计离散化,其优化能力并不如 LID。

**表 2　在 IoU=0.7 的 Car 下的消融实验结果**

**Table 2　Experimental results of ablation under Car with IOU=0.7**　(%)

| 实验 | Basebone | SAM | C-ASPP | D | SID | LID | Car (IoU=0.7) | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Easy | Mod. | Hard |
| 1 | √ | | | | | | 20.17 | 13.41 | 11.86 |
| 2 | √ | √ | | | | | 21.02 | 13.69 | 12.25 |
| 3 | √ | √ | √ | | | | 21.61 | 14.63 | 12.19 |
| 4 | √ | √ | √ | √ | | | 22.64 | 15.37 | 13.13 |
| 5 | √ | √ | √ | √ | | | 23.24 | 16.11 | 13.86 |
| 6 | √ | √ | √ | √ | √ | | 23.99 | 16.26 | 14.14 |
| 本文 | √ | √ | √ | √ | | √ | 24.57 | 16.78 | 14.41 |

2)P-BEV 消融实验

透视 BEV 能够有效地消除特征采样带来的干扰,来提高检测精度。将 BEV 与 P-BEV 进行对比实验,其结果如表 3 所示。通过数据分析,在利用 P-BEV 代替普通 BEV 作为最终检测输入,能够有效地提升网络检测精度,其在 Car(IoU=0.7)实验中,Easy、Mod 以及 Hard 均有提升。

**表 3　P-BEV 消融实验**

**Table 3　Perspective BEV ablation experiment**　(%)

| 实验 | Car (IoU=0.7) | | |
|---|---|---|---|
| | Easy | Mod. | Hard |
| BEV | 20.17 | 13.41 | 11.86 |
| P-BEV | 21.72 | 14.26 | 12.65 |

## 2.3　实验结果可视化

为验证 CDCN-3D 网络实际的检测效果,在各种不同场景下使用 RGB 图片进行 3D 目标检测任务。结果表明,CDCN-3D 网络具有较高的精度和鲁棒性,能够准确检测不同形状、大小、颜色和光照条件下的各种目标,包括汽车、行人、自行车等。这进一步证明了 CDCN-3D 网络的有效性和实用性,有望在实际应用中发挥重要作用。

1)简单场景检测效果

为验证本文算法的可行性,在一些简单场景下,如所有目标均可视、目标之间无遮挡等场景下对网络进行测试,如图 8 所示,分别为:Ⅰ):少量行人;Ⅱ):少量汽车;Ⅲ):汽车和骑行者;Ⅳ):数量多的行人和骑行者;Ⅵ):数量多的汽车;Ⅴ):行人、骑行者以及各种汽车。在图中所展示的简单场景中,通过对一些相对简单的场景目标进行检测来验证 CDCN-3D 网络检测可行性。图中每个场景都输出其二维检测结果、深度估计结果、透视 BEV 以及最终的 3D 检测结果。通过分析实验结果图,网络面对简单场景时,如Ⅰ中单目标以及Ⅵ独立多目标,网络能够较好地完成检测任务。

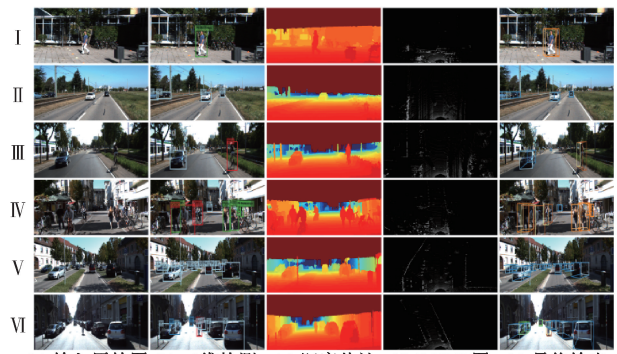


图 8　CDCN-3D 网络在简单场景下的测试结果

Fig. 8　Test results of CDCN-3D network in a simple scenario

2)复杂场景检测效果

为验证本文算法的精确性,在一些复杂场景如目标信息缺失、目标之间无遮挡等场景下对网络进行测试,如图 9 所示,场景分别为:Ⅰ):汽车被建筑物遮挡;Ⅱ):汽车之间存在遮挡;Ⅲ):汽车被行人遮挡;Ⅳ):十字路口复杂路况;Ⅵ):汽车与骑行者以及存在行人;Ⅴ):行人、骑行者以及各种汽车相互遮挡。汽车和骑行车并行以及行人和骑行车相互交错的情况下,在最终的检测图可以发现,网络能够准确地识别不同类别检测目标,且网络面对信息缺失目标时,仍能准确地完成检测任务。

3)误差分析

实验结果证明,大部分情况下,都能有效地完成检测任务,实现单目 3D 目标检测。但可以发现,在相同大小物体存在大面积遮挡的情况下,如图 10Ⅰ中,会出现个别目标被漏检;以及在多类复杂目标时如图 10Ⅱ和Ⅲ,其 3D BOX 定位会存在一定偏差。这是由于当目标信息

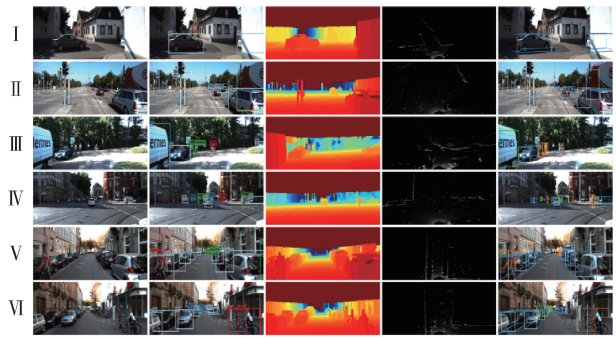


图 9　CDCN-3D 网络在复杂场景下的测试结果

Fig. 9　Test results of CDCN-3D network in complex scenes

严重缺失甚至特征不可见时,周围目标对其产生信息干扰时网络无法学习获取其深度信息,会错误地将其忽略从而造成漏检误检甚至标定错误的情况发生。


图 10　误差结果图

Fig. 10　Error result chart

## 3　结　论

本文针对单目 3D 目标检测深度信息估算难和特征融合时存在深度信息丢失的问题,提出一种用于单目 3D 对象检测的端到端联合多注意力和深度估计检测网络模型 CDCN-3D,在编码器解码器的框架上,引入 C-ASPP 模块,从而实现自适应提取编码器特征信息,生成高质量深度图,同时利用精确且高效率的 P-BEV 将得到的物体三维信息投影到二维平面。为了能够快速完成单目 3D 物体检测,提出简单有效地网络结构以端到端学习的方式联合深度估计和三维目标检测。

研究表明,CDCN-3D 网络在 KITTI 数据集能对各种场景下完成检测任务,且评价指标高于其他算法。但在环境复杂,如相同物体存在重叠的情况下,会出现漏检问题,需要在后续研究中继续优化算法,提升检测精度,开发出检测速度更快、精度更高的网络。

## 参考文献

[1] 伍锡如,邱涛涛,王耀南. 改进 Mask R-CNN 的交通场景多目标快速检测与分割[J]. 仪器仪表学报,2021,42(7):242-249.
WU X R, QIU T T, WANG Y N. Improved Mask R-CNN multi-target fast detection and segmentation of traffic scenes [J]. Chinese Journal of Scientific Instrument, 2021,42(7):242-249.

[2] 郑少武,李巍华,胡坚耀. 基于激光点云与图像信息融合的交通环境车辆检测[J]. 仪器仪表学报,2019,40(12):143-151.
ZHENG SH W, LI W H, HU J Y. Vehicle detection in traffic environment based on laser point cloud and image information fusion [J]. Chinese Journal of Scientific Instrument, 2019,40(12):143-151.

[3] ZIMMER W, ERCELIK E, ZHOU X, et al. A survey of robust 3D object detection methods in point clouds[J]. arXiv preprint arXiv:2204.00106, 2022.

[4] THILAKANAYAKE T, HERATH N, LIYANAGE M. Development of a stereo vision-based pick and place system for robotic manipulators [J]. Instrumentation, 2021,8(2):1-13.

[5] QIN Z, WANG J, LU Y. Triangulation learning network: From monocular to stereo 3D object detection [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019: 7615-7623.

[6] 张浩,左杭,刘宝华. 视觉与二维激光雷达的目标检测方法[J]. 电子测量与仪器学报,2022,36(3):79-86.
ZHANG H, ZUO H, LIU B H. Target detection method of vision and two-dimensional Lidar [J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(3):79-86.

[7] 张艳邦,张芬. 融合纹理和颜色特征的显著目标检测[J]. 计算机与数字工程,2021,49(9):1793-1798.
ZHANG Y B, ZHANG F. Detection of salient objects with texture and color features [J]. Computer and Digital Engineering, 2021,49(9):1793-1798.

[8] LI B, OU Y W, SHENG L, et al. GS3D: An efficient 3D object detection framework for autonomous driving[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 1019-1028.

[9] LIU Z, ZHOU D, LU F, et al. AutoShape: Real-time shape-aware monocular 3D object detection [C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 15641-15650.

[10] CHABOT F, CHAOUCH M, RABARISOA J, et al. Deep

MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, 2017: 1827-1836.

[11] XU B, CHEN Z. Multi-level fusion based 3D object detection from monocular images[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 2345-2353.

[12] SUN Y, LI Z, WANG L, et al. Automatic detection of vehicle targets based on centernet model[C]. 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE). IEEE, 2021: 375-378.

[13] BRAZIL G, LIU X. M3d-rpn: Monocular 3D region proposal network for object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9287-9296.

[14] ZHAO C, SUN Q, ZHANG C, et al. Monocular depth estimation based on deep learning: An overview[J]. Science China Technological Sciences, 2020, 63(9): 1612-1627.

[15] LI Y, GE Z, YU G, et al. Bevdepth: Acquisition of reliable depth for multi-view 3D object detection[J]. arXiv preprint arXiv:2206.10092, 2022.

[16] ZHAO C, SUN Q, ZHANG C, et al. Monocular depth estimation based on deep learning: An overview[J]. Science China Technological Sciences, 2020, 63(9): 1612-1627.

[17] WANG Y, CHAO W L, GARG D, et al. Pseudo-Lidar from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 8445-8453.

[18] 张启超,乔战伟. 基于单目视觉的物体绝对深度测量研究[J]. 电子测量技术,2020,43(20):74-78.
ZHANG Q CH, QIAO ZH W. Research on absolute depth measurement of objects based on monocular vision[J]. Electronic Measurement Technology, 2020, 43(20): 74-78.

[19] 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[J]. 计算机学报, 2022, 45(6): 1276-1307.
JIANG J J, LI ZH Y, LIU X M. Overview of monocular depth estimation methods based on deep learning[J]. Chinese Journal of Computers, 2022,45(6):1276-1307.

[20] SAXENA A, SCHULTE J, NG A Y. Depth estimation using monocular and stereo cues[C]. IJCAI. 2007, 7: 2197-2203.

[21] LIU F, SHEN C, LIN G. Deep convolutional neural fields for depth estimation from a single image[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 5162-5170.

[22] WANG P, SHEN X, LIN Z, et al. Towards unified depth and semantic prediction from a single image[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 2800-2809.

[23] JONAH P, SANJA F. Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D[C]. ECCV, 2020.

[24] RODDICK T, KENDALL A, CIPOLLA R. Orthographic feature transform for monocular 3D object detection[J]. arXiv preprint arXiv:1811.08188, 2018.

[25] READING C, HARAKEH A, CHAE J, et al. Categorical depth distribution network for monocular 3D object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 8555-8564.

[26] WU H Z, YAN L F, LIU X Q, et al. The feature ambiguity mitigate operator model helps improve bone fracture detection on X-ray radiograph[J]. Scientific Reports, 2021, 11(1): 1-10.

[27] 李瑞龙,吴川,朱明. 体素化点云场景下的三维目标检测[J]. 液晶与显示,2022,37(10):1355-1363.
LI R L, WU CH, ZHU M. 3D target detection in voxelized point cloud scene[J]. Chinese Journal of Liquid Crystals and Displays, 2022,37(10):1355-1363.

## 作者简介

**郑自立**,2020年于南通大学获得学士学位,现于西安工程大学攻读硕士学位,主要研究方向为机器视觉、3D目标检测。
E-mail: zzl513x@163.com

**Zheng Zili** received his B. Sc. degree from Nantong University in 2020. Now he is a M. Sc. candidate at Xi′an Engineering University. His main research interests include machine vision and 3D object detection.

**徐健**(通信作者),1986年于西安工程大学获得学士学位,现为西安工程大学教授,主要研究方向为机器视觉、目标检测、目标跟踪。
E-mail: xu0910@sian.com

**Xu Jian**(Corresponding author) received his B. Sc. degree from Xi′an Engineering University in 1986. Now he is a professor at Xi′an Engineering University. His main research interests include machine vision, target detection and target tracking.