第5章　环境感知与识别

5.1　环境感知与识别概述

环境感知对象主要包括行驶路径、周边物体、驾驶状态、驾驶环境。其中行驶路径主要包括结构化道路和非结构化道路两大块，其中结构化道路包括车道线、道路边缘、道路隔离物、恶劣路况的识别，非结构化道路包括可行驶路径的确认和前方路面环境的识别。周边物体主要包括车辆、行人、地面上可能影响车辆通过性、安全性的其他各种移动或静止障碍物的识别及各种交通标志的识别。本章重点讨论行驶路径部分的车道线检测以及周边物体中的障碍物检测、红绿灯检测。

环境感知与识别传感器系统通常采用摄像头、激光雷达、毫米波雷达等多种车载传感器来感知环境。就三种传感器的应用特点来讲，摄像头和激光雷达都可用于进行车道线检测。对红绿灯的识别，主要还是用摄像头来完成。而对障碍物的识别，摄像头可以通过深度学习把障碍物进行细致分类，激光雷达只能分一些大类，但能完成对物体距离的准确定位；毫米波雷达则完成障碍物运动速度、方位等识别。

5.2　障碍物检测

车辆行驶道路上的障碍物检测是无人驾驶汽车环境感知模块中的重要组成部分。准确的障碍物检测决定着无人驾驶汽车行驶的安全性。目前障碍物检测技术主要包括以下三种方法：

（1）基于图像的障碍物检测。

（2）基于激光雷达的障碍物检测。

（3）基于视觉和激光雷达融合的障碍物检测。

5.2.1　基于图像的障碍物检测

基于图像的障碍物检测算法已经发展得较成熟了，大致可以分为一阶段检测算法和二阶段检测算法。一阶段检测算法有 YOLO 和 SSD

等，二阶段检测算法则主要是 RCNN 这一流派。当前的二阶段检测算法大多是在 Faster RCNN 基础上的改进。两种检测算法相比，一阶段算法的速度是快于二阶段算法的，而在准确度上，二阶段算法更胜一筹。

1. 基于二维图像的障碍物检测

1) YOLO 系列障碍物检测

YOLO(You Only Look Once)是将物体检测作为回归问题求解的一种一阶段检测算法。它基于一个单独的端到端网络，完成从原始图像的输入到物体位置和类别的输出。从网络设计上，YOLO 与 RCNN、Fast RCNN 及 Faster RCNN 的区别如下：

(1) YOLO 训练和检测均是在一个单独网络中进行，没有显式地求取区域候选框的过程，这是它相比基于候选框方法的优势。而 RCNN/Fast RCNN 采用分离的模块(独立于网络之外的选择性搜索方法)求取候选框(可能会包含物体的矩形区域)，训练过程因此也是分成多个模块进行。Faster RCNN 使用 RPN(Region Proposal Network)卷积网络替代 RCNN/Fast RCNN 的选择性搜索模块，将 RPN 集成到 Fast RCNN 检测网络中，得到一个统一的检测网络。尽管 RPN 与 Fast RCNN 共享卷积层，但是在模型训练过程中，需要反复训练 RPN 网络和 Fast RCNN 网络(这两个网络核心卷积层是参数共享的)。

(2) YOLO 输入图像经过一次推理，便能得到图像中所有物体的位置和其所属类别及相应的置信概率。而 RCNN/Fast RCNN/Faster RCNN 将检测结果分为两部分求解：物体类别(分类问题)和物体位置，即标注框(bounding box)。

接下来介绍一下 YOLO 系列障碍物检测核心思想。

(1) 网络定义。

YOLO 检测网络包括 24 个卷积层和 2 个全连接层。其中，卷积层用来提取图像特征，全连接层用来预测图像位置和类别概率值。YOLO 网络借鉴了 GoogleNet 分类网络结构。不同的是，YOLO 未使用 Inception 模块，而是使用 1×1 卷积层(此处 1×1 卷积层的存在是为了跨通道信息整合)和 3×3 卷积层简单替代。

YOLO 论文中，作者还给出一个更轻快的检测网络 fast YOLO，它只有 9 个卷积层和 2 个全连接层。使用 Titan X GPU，fast YOLO 可以达到 155f/s 的检测速度，但是 mAP 值也从 YOLO 的 63.4%降到了 52.7%，但却仍然远高于以往的实时物体检测方法(DPM)的 mAP 值。

(2) 输出表达(representation)定义。

本部分给出 YOLO 全连接输出层的定义。

YOLO 将输入图像分成 $S\times S$ 个格子，每个格子负责检测“落入”该格子的物体。若某个物体的中心位置的坐标落入到某个格子，那么这个格子就负责检测出这个物体。如图 5-1 所示，图中物体狗的中心点(红色原点)落入第 5 行第 2 列的格子内，所以这个格子负责预测图像中的物体狗。

每个格子输出 B 个标注框(包含物体的矩形区域)信息，以及 C 个物体属于某种类别的概率信息。标注框信息包含 5 个数据值，分别是 x、y、w、h 和 confidence。其中 x、y 是指当前格子预测得到的物体的标注框的中心位置的坐标。w、h 是标注框的宽度和高度。注意，实际训练过程中，w 和 h 的值使用图像的宽度和高度进行归一化到[0,1]区间内；x、y

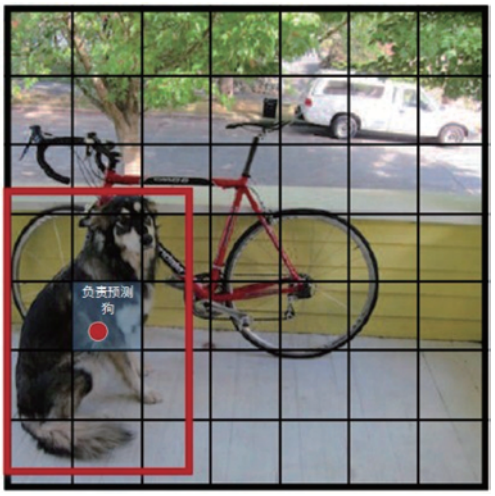


图 5-1 YOLO 检测狗效果图

是标注框中心位置相对于当前格子位置的偏移值，并且被归一化到[0,1]。confidence 反映当前标注框是否包含物体以及物体位置的准确性，计算方式如下：

$$\text{confidence} = P(\text{object})$$

其中，若标注框包含物体，则 $P(\text{object})=1$；否则 $P(\text{object})=0$。

因此，YOLO 网络最终的全连接层的输出维度是 $S\times S\times(B\times 5+C)$。YOLO 论文中，作者训练采用的输入图像分辨率是 448×448 像素，$S=7$，$B=2$；采用 VOC 20 类标注物体作为训练数据，$C=20$。因此输出向量为 $7\times 7\times(20+2\times 5)=1470$ 维。

注：

① IOU(Intersection Over Union)为预测标注框与物体真实区域的交集面积(以像素为单位，用真实区域的像素面积归一化到[0,1]区间)。

② 由于输出层为全连接层，因此在检测时，YOLO 训练模型只支持与训练图像相同的输入分辨率。

③ 虽然每个格子可以预测 B 个标注框，但是最终只选择 IOU 最高的标注框作为物体检测输出，即每个格子最多只预测出一个物体。当物体占画面比例较小，如图像中包含畜群或鸟群时，每个格子包含多个物体，但却只能检测出其中一个。这是 YOLO 方法的一个缺陷。

(3) 损失(loss)函数定义。

YOLO 使用均方和误差作为 loss 函数来优化模型参数，即网络输出的 $S\times S\times(B\times 5+C)$ 维向量与真实图像的对应 $S\times S\times(B\times 5+C)$ 维向量的均方和误差。如下式所示，其中，coordError、iouError 和 classError 分别代表预测数据与标定数据之间的坐标误差、IOU 误差和分类误差。

$$\text{loss}=\sum_{i=0}^{S^2}\text{coordError}+\text{iouError}+\text{classError}$$

YOLO 对上式 loss 的计算进行了如下修正：

(1) 位置相关误差(坐标、IOU)与分类误差对网络 loss 的贡献值是不同的，因此 YOLO

在计算 loss 时，使用 $\lambda_{coord}=5$ 修正 coordError。

(2) 在计算 IOU 误差时，包含物体的格子与不包含物体的格子，二者的 IOU 误差对网络 loss 的贡献值是不同的。若采用相同的权值，那么不包含物体的格子的 confidence 值近似为 0，变相放大了包含物体的格子的 confidence 误差在计算网络参数梯度时的影响。为解决这个问题，YOLO 使用 $\lambda_{noobj}=0.5$ 修正 iouError(此处的“包含”是指存在一个物体，它的中心坐标落入到格子内)。

(3) 对于相等的误差值，大物体误差对检测的影响应小于小物体误差对检测的影响。这是因为，相同的位置偏差占大物体的比例远小于同等偏差占小物体的比例。YOLO 将物体大小的信息项(w 和 h)进行求平方根来改进这个问题(注：这个方法并不能完全解决这个问题)。

注：

① YOLO 方法模型训练依赖于物体识别标注数据，因此，对于非常规的物体形状或比例，YOLO 的检测效果并不理想。

② YOLO 采用了多个下采样层，网络学到的物体特征并不精细，因此也会影响检测效果。

③ YOLO loss 函数中，大物体 IOU 误差和小物体 IOU 误差对网络训练中 loss 贡献值接近(虽然采用求平方根方式，但没有根本解决问题)。因此，对于小物体，小的 IOU 误差也会对网络优化过程造成很大的影响，从而降低了物体检测的定位准确性。

2) SSD 障碍物检测

SSD(Single Shot Multibox Detector)也是一种单一阶段检测算法，只需要用到图像一次，无须先产生候选框再进行分类和回归，而是直接在图像中不同位置进行边界框的采样，然后使用卷积层进行特征提取后直接进行分类和回归。相比基于候选框的方法，SSD 极大地提高了的检测速度。

接下来介绍一下 SSD 检测的主要设计理念。

(1) 使用不同尺度下的特征图进行检测。

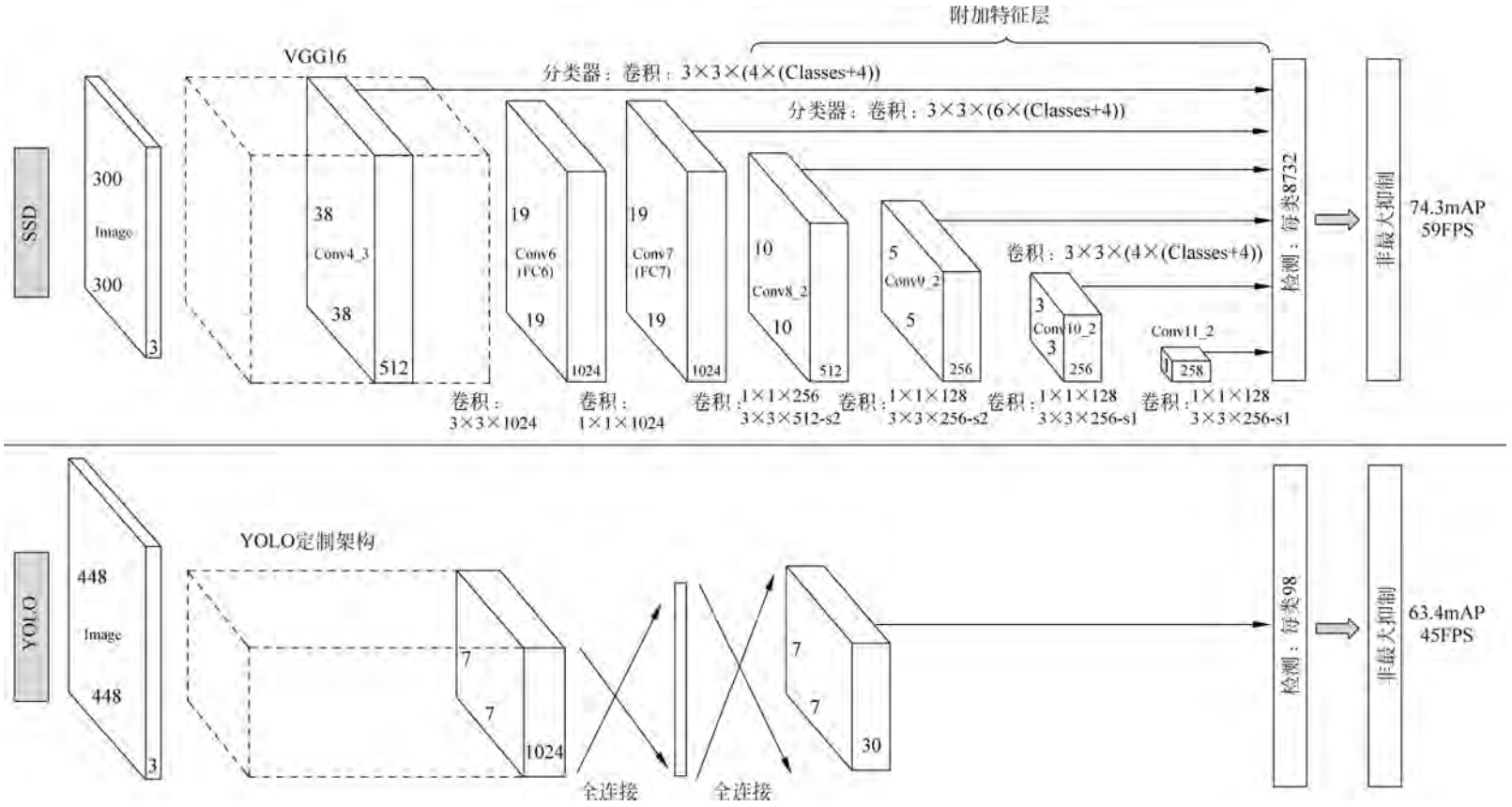
SSD 通过提取不同尺度下的特征图来做检测，其网络结构与 YOLO 的比较如图 5-2 所示，它使用了六种不同尺寸的特征图来进行检测。在卷积神经网络中，较低层级的特征图尺寸较大，在这种特征图上的候选框在原图上的覆盖范围较小；较高层级的特征图的尺寸较小，而其候选框在原图上的覆盖范围大。如图 5-3(b)和图 5-3(c)所示，尺寸为 8×8 的特征图划分了更多单元格，但是每个单元格在原图中所占范围较小。通过对多个层级上的候选框进行匹配，有利于更精确地找到与不同尺寸目标最匹配的边界框。

(2) 采用卷积层做检测。

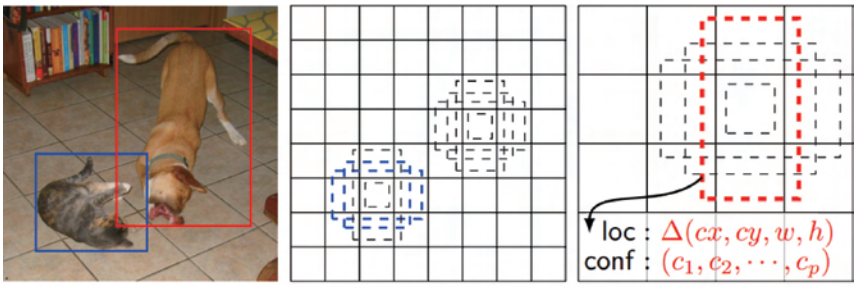
与 YOLO 在全连接层之后做检测不同的是，SSD 直接采用卷积对不同特征图进行特征提取。如图 5-2 所示，对于尺寸为 $m\times n$，维数为 p 的特征图，SSD 使用 $3\times3\times p$ 的卷积核来进行卷积。其输出一种为用于分类的置信度分数，另一种为用于回归的位移量。另外，卷积层还通过候选框层生成候选框坐标，每层产生的候选框个数是一定的。

(3) 采用不同尺度和纵横比的候选框。

SSD 借鉴了与 Faster RCNN 类似的候选框思想，在每个单元格设置不同尺度和纵横比的默认框，如图 5-3 所示，每个单元格设定有四种不同尺度的默认框。对于一个尺寸为 $m\times n$



■图 5-2　SSD 与 YOLO 网络结构比较。图片来源于 Liu，et al.，2016



(a) 带真实标注的图像　(b) 8×8特征图　(c) 4×4特征图

■图 5-3　SSD 基本框架。图片来源于 Liu,et al.，2016

的特征图，假定每个单元格有 k 个默认框，则该特征图共有 $m\times n\times k$ 个默认框，如图 5-2 中，该网络共生成了 8732(38×38×4+19×19×6+10×10×6+5×5×6+3×3×4+1×1×4)个默认框。

对于每个默认框，SSD 的预测值主要有两方面：分类的置信度和边界框的回归值。在分类中，需要注意的是 SSD 把背景也单独作为一类，如在 VOC 数据集上，SSD 的每个默认框会输出 21 类置信度，其中 20 类为 VOC 的目标种类。边界框的回归值与 Faster RCNN 类似，实际上是预测真实边界框 g 相对于默认框 d 的中心(cx，cy)和宽(w)、高(h)的转换量，预测值的真实值的计算方式为

$$\hat{g}^{cx}=(g^{cx}-d^{cx})/d^{w},\quad \hat{g}^{cy}=(g^{cy}-d^{cy})/d^{h}$$

$$\hat{g}^{w}=\log\left(\frac{g^{w}}{d^{w}}\right),\quad \hat{g}^{h}=\log\left(\frac{g^{h}}{d^{h}}\right)$$

因此，假定该数据集有 c 种目标，则每个默认框需要预测 $c+1$ 个类别概率和 4 个坐标相关的转换量。

在训练过程中，SSD 会在开始阶段将这些默认框与真实框进行匹配，如在如图 5-3 中，对猫和狗分别采用适合它们形状的默认框，蓝色框匹配到了猫的真实框，红色框匹配到了狗的真实框。在匹配过程中，首先对于每个真实框，找与其 IOU 最大的默认框进行匹配；然后，对于剩余的默认框，若其与某一真实框的 IOU 大于设定阈值(文中为 0.5)，SSD 也会将它们进行匹配。也就是说，一个真实框可能会与多个默认框匹配。这种匹配方法简化了网络的学习问题，使得网络可以在多个框中选择预测分数最高的，而不是只能用重合度最大的框来做预测。

另外，与真实框匹配的默认框记为正样本，未匹配的记为负样本，显然这样产生的负样本要远远多于正样本。为了保证正负样本比例尽量平衡，SSD 将负样本按置信度误差排序，并选择排名靠前的，使得负样本与正样本的比例约为 3∶1。

(4) 损失(loss)函数定义。

SSD 的损失函数由位置误差(localization loss，loc)和置信度误差(confidence loss，conf)组成。令 $x_{ij}^{p}=\{1,0\}$ 表示第 i 个默认框是否与第 j 个真实框匹配，N 为匹配的默认框总数，c 为类别置信度预测值，g 为真实边界框，l 为预测框，则总的损失函数为

$$L(x,c,l,g)=\frac{1}{N}(L_{\text{conf}}(x,c)+\alpha L_{\text{loc}}(x,l,g))$$

对于位置误差 L_{loc}，采用了 Smooth L1 loss(平滑的 L1 损失)：

$$L_{\text{loc}}(x,l,g)=\sum_{i\in \text{Pos}}^{N}\sum_{m\in\{cx,cy,w,h\}}x_{ij}^{k}\,\text{smooth}_{\text{L1}}(l_i^m-\hat{g}_j^m)$$

其中，真实转换值 $\hat{g}$ 的计算方式在(3)中已给出。

对于置信度误差 L_{conf}，采用了 softmax loss：

$$L_{\text{conf}}(x,c)=-\sum_{i\in \text{Pos}}^{N}x_{ij}^{p}\log(\hat{c}_i^p)-\sum_{i\in \text{Neg}}\log(\hat{c}_i^p)$$

$$\hat{c}_i^p=\frac{\exp(c_i^p)}{\sum_p\exp(c_i^p)}$$

总误差函数中的权重系数 α 通过交叉验证设置。

3) Faster RCNN 障碍物检测

经过 RCNN 和 Fast RCNN 的积淀，Ross B. Girshick 等人在 2016 年提出了新目标检测框——Faster RCNN，不同于 YOLO，它是一种二阶段的检测算法。RCNN 系列算法如图 5-4 所示，RCNN 以及 Fast RCNN 都无法做到端到端的训练，Faster RCNN 则将体征提取模块、候选框生成模块以及边框回归和目标分类模块都整合在了一个网络中，使得综合性能有较大提高，在检测速度方面尤为明显。

图 5-4 RCNN 系列算法

(1) 特征提取模块。

Faster RCNN 特征提取网络使用的 VGG16 是牛津大学计算机视觉组和 Deepmind 公司的研究人员一起研发的深度卷积神经网络。他们探索卷积神经网络中的深度、宽度和性能之间的关系，通过反复堆叠 3×3 卷积和 2×2 的最大值池化，成功构建一个 16 层的网络。输入图像大小是 3×224×224，输出特征则为 51×39×256。

(2) 候选框生成模块(RPN)。

经典的检测方法生成检测框都非常耗时，如 Opencv 的 Adaboost 方法使用滑动窗口和图像金字塔生成检测框；或如 RCNN 使用 SS(Selective Search)方法生成检测框。而 Faster RCNN 则抛弃了传统的滑动窗口和 SS 方法，直接使用 Region Proposal Networks(RPN)生成检测框，这也是 Faster RCNN 的巨大优势，能极大提升检测框的生成速度。

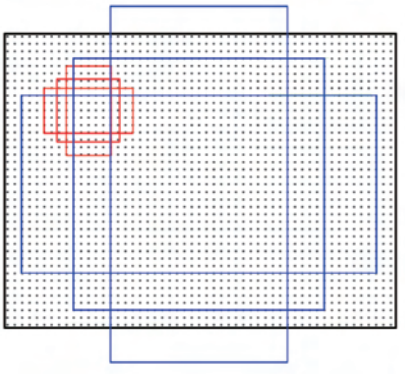


图 5-5 anchor 生成图

特征可以看作一个尺度 51×39 的 256 通道图像，对于该图像的每一个位置(如图 5-5 所示)，考虑

9 个可能的候选窗口：三种面积$\{128^2,256^2,512^2\}$×三种比例$\{1:1,1:2,2:1\}$，这些候选窗口称为锚点(anchor)。

RPN 实际分为两个组件：一个组件通过 softmax 分类 anchor 获得前景和背景；另一个组件用于计算对 anchor 的边框偏移量，以获得精确的候选框。

(3) 边框回归和目标分类模块。

在通过 RPN 得到候选框之后，使用 ROI pooling 将每个候选框所对应的特征都转换成 7×7 的大小。再将每个候选框的特征输入到边框回归和目标分类模块中，得到每个候选框的类别，类别数是 $n+1$，n 是 n 种障碍物类别，1 则是背景。对于那些非背景的目标框，则会进行边框修正。

在 Faster RCNN 中主要是两个损失：一个是分类损失；另一个是标注框回归损失。对于分类损失来说就是一个简单的交叉熵，而标注框回归损失采用的是平滑的 L1 损失。

$$\text{smooth}_{L1}(x)=\begin{cases}0.5x^2\times\dfrac{1}{\sigma^2} & |x|<\dfrac{1}{\sigma^2}\\ |x|-0.5 & \text{其他}\end{cases}$$

Faster RCNN 目标检测算法是一种主流的二阶段算法，它在精度和速度上有着不错的表现。一些基于 Faster RCNN 的改进算法，如特征金字塔、可形变卷积以及级联 RCNN 等，都能取得了目前最好的检测效果。

2. 基于图像的三维障碍物检测

尽管 Faster RCNN、YOLO 等算法能够准确地检测出障碍物在图像中的位置，但是现实场景是三维的，物体都是三维形状的，大部分应用都需要有目标物体的三维的长宽高、空间信息、朝向信息偏转角等。例如在图 5-6 中，在自动驾驶场景下，需要从图像中提供目标物体长宽高、空间信息、朝向信息偏转角等指标，鸟瞰投影的信息对于后续自动驾驶场景中的路径规划和控制具有至关重要的作用。

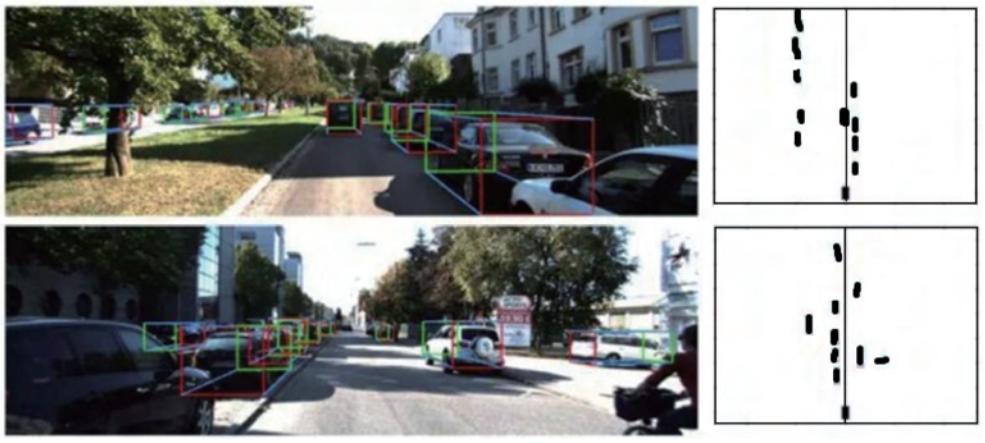


图 5-6　三维障碍物检测以及鸟瞰效果图

目前三维目标检测正处于高速发展时期，主要是综合利用单目摄像头、双目摄像头、多线激光雷达来进行三维目标检测。从目前成本上讲，激光雷达＞双目摄像头＞单目摄像头，从目前的准确率上讲，激光雷达＞双目摄像头＞单目摄像头。但是随着激光雷达的不断产

业化发展，成本在不断降低，也出现一些使用单目摄像头加线数较少的激光雷达进行的技术方案。

如图 5-7 所示，以开源的 Apollo 为例，Apollo 中使用的 YOLO 三维，在 Apollo 中通过一个多任务网络来进行车道线和场景中目标物体检测。其中的 Encoder 模块是 YOLO 的 Darknet，在原始 Darknet 基础上加入了更深的卷积层，同时添加反卷积层，捕捉更丰富的图像上下文信息。高分辨多通道特征图可以捕捉图像细节；深层低分辨率多通道特征图可以编码更多图像上下文信息。YOLO 的 Darknet 采用和 FPN(Feature Pyramid Network)类似的连接，更好地融合了图像的细节和整体信息。Decoder 分为两个部分：一部分为语义分割，用于车道线检测；另一部分为物体检测，物体检测部分基于 YOLO，同时还会输出物体的方向等三维信息。

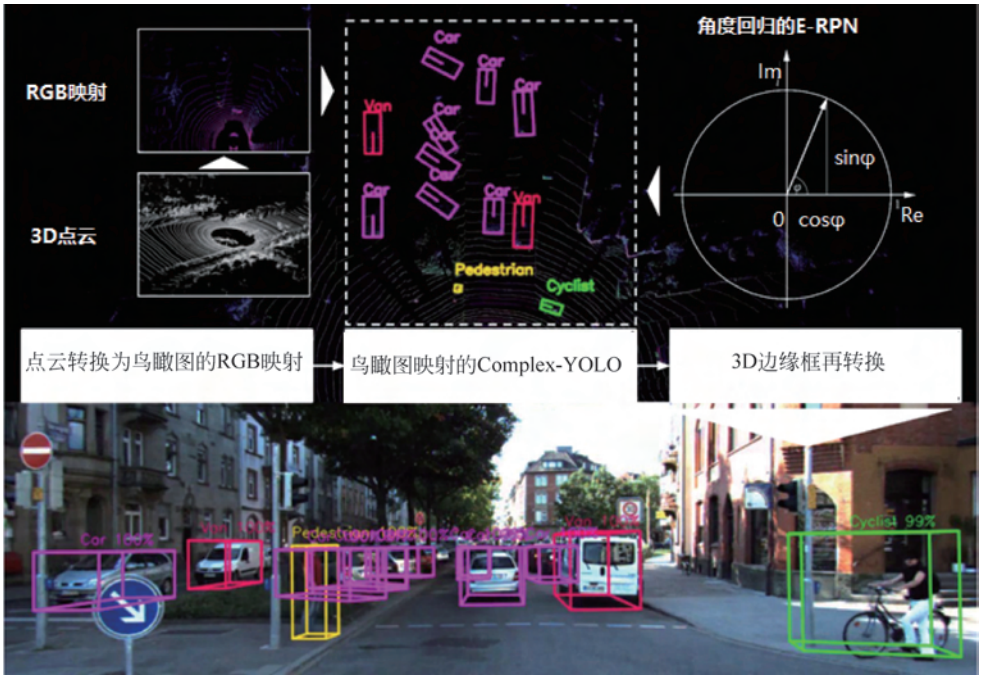


图 5-7 YOLO 三维。图片来源于 Simon，el at.，2018

利用地面平行假设，来降低所需要预测的三维参数。①假设三维障碍物只沿着垂直地面的坐标轴有旋转，而另外两个方向并未出现旋转。障碍物中心高度和摄像头高度相当，所以可以简化认为障碍物的 $Z=0$；②可以利用成熟的二维障碍物检测算法，准确预测出图像上二维障碍物框(以像素为单位)；③对三维障碍物里的 6 维(中心点坐标，长、宽、高)描述，可以选择训练神经网络来预测方差较小的参数。

5.2.2 基于激光雷达的障碍物检测

近年来，利用雷达或图像与雷达融合的障碍物检测方法逐渐被普及，扫描距离远、精度高的激光雷达便是非常理想的障碍物检测工具之一。本节将介绍基于激光雷达的障碍物检测的几种方法。

1. 基于几何特征和网格

几何特征包括直线、圆和矩形等。基于几何特征的方法首先对激光雷达的数据进行处理，采用聚类算法将数据聚类并与障碍物的几何特征进行对比，对障碍物进行检测和分类。利用几何特征的方法在无人驾驶方面较为常见。

为了提高对不同点云数据检测的可靠性，基于几何特征的方法也可以与光谱特征结合，将几何和影像特征综合考虑，从多个维度对障碍物进行识别，同时引入权重系数来反映不同的特征对识别的影响。

对于非结构化的道路，障碍物的形状复杂，较难用几何形状去描述，此时需要用基于网格的方法来识别此类障碍物。该方法将激光雷达的数据投影到网格地图中，然后利用无向图相关方法对点云数据进行处理。网格的大小和结构可以自定义，用网格分布图像来表示障碍物，分辨率越高的网格，表示的障碍物越复杂，但同时需要较高的计算复杂度和内存。

2. VoxelNet 障碍物检测

为了方便进行障碍物检测，激光雷达数据需要一定的人力对数据进行整理，对于距离较远的物体，激光雷达扫描出的物体轮廓信息在网络进行识别时效果可能并不理想，为此需要投入更多的人工流程去处理激光雷达数据。为了解决这一难点，减少人力，在 VoxelNet 的研究中，消除了对点云进行手动提取特征的过程，并提出了统一的端到端的三维检测网络。

VoxelNet 将原始点云作为输入，但由于现有激光雷达返回的点云数据包含的坐标点数量较大，多为数万到数十万，对计算量和内存的需求过大，VoxelNet 除了减少人力标注外，也着重解决了如何让网络高效处理更多的激光点云数据。如图 5-8 所示，VoxelNet 主要由三个模块组成：特征学习网络、中间卷积层和区域建议网络。

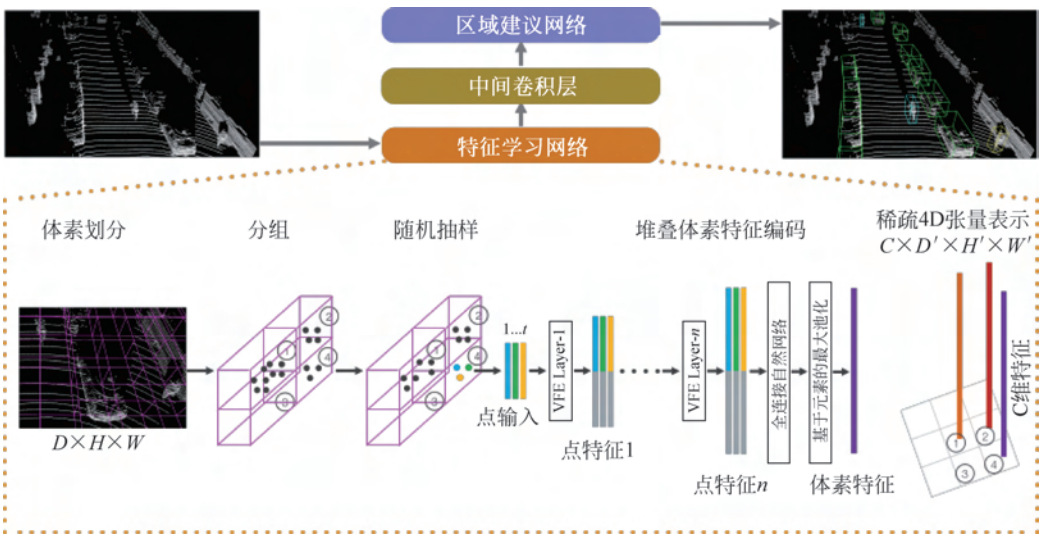


图 5-8 VoxelNet 主要构成模块。图片来源于 Zhou Y，Tuzel O，2017

特征学习网络主要将点云划分为体素 Voxel 形式，通过 VFE 层提取特征，得到体素级的特征向量，包括以下几个步骤。

（1）体素划分：给定输入点云，需要将空间划分成均匀的体素 Voxel，这里假设点云对

应的 x、y、z 坐标信息分别代表 W、H、D 的三维空间，V_w、V_h、V_d 分别为每个体素的大小，由此可以得到 $D' \times H' \times W'$ 个体素网格，这里 $D'=D/V_d$、$H'=H/V_h$、$W'=W/V_h$。

(2) 分组：将点云根据空间位置划分到相对应的体素中，但由于距离、遮挡、物体相对姿态和采样不均匀等原因，激光雷达获取的点云在空间中的分布不均匀，所以分组时会造成不同的体素中包含的点云数量不同。

(3) 随机抽样：由于激光雷达获取的点云数据较多，直接进行处理计算和内存的负荷较大，并且由于采样不均匀等原因可能会对结果造成影响。所以 VoxelNet 采用了随机抽样的策略，具体方法是随机从拥有多于 T 个点的体素中随机采样 T 个点云。随机抽样的策略从一定程度上减少了体素之间点云分布不平衡，减少了采样的偏移，更有利于训练，同时节省了计算量。整个过程如图 5-9 所示。

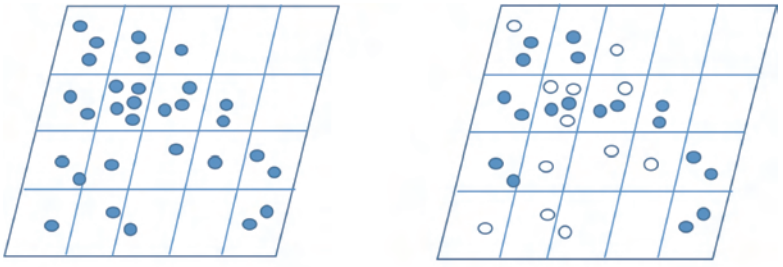


图 5-9 体素划分与分组和随机抽样(深色点云代表 $T=2$ 时采样点云)

(4) 堆叠体素特征编码：主要通过级联的 VFE 层实现基于点的点特征和局部特征的融合，以第一层 VFE 层为例，主要流程如图 5-10 所示。首先对每个网格中的点云进行去中心化，得到的每个点的 VFE 层的输入；每个点经过包含 ReLU 函数和 BN(Batch Normal)运算的全连接网络，得到点特征；对每个点特征都进行最大池化运算，得到局部聚合特征；最终将点特征和局部聚合特征进行结合运算，得到最后的特征向量，对每个体素进行处理，得到特征提取层的输出。

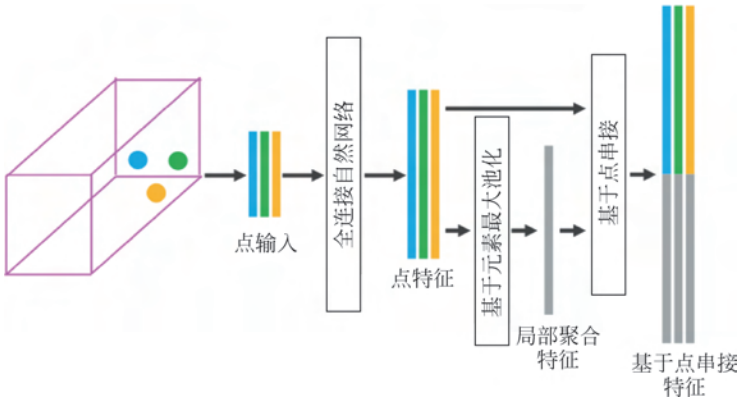


图 5-10 堆叠体素特征 VFE 层流程图。图片来源于 Zhou Y，Tuzel O，2017

(5) 稀疏张量表示：在体素划分时，许多体素(超过 90%)是空的，所以只需要对非空体素进行 VFE 处理，将其表示为稀疏张量，从而有效节省资源。

中间卷积层负责将特征向量进行三维卷积，提取特征，获取全局特征。

区域建议网络(RPN)将特征进行整合，输出预测概率，给出预测结果等。VoxelNet 的 RPN 设计与前面提到的 Faster RCNN 的设计类似。损失函数的设定同样也是只关注置信度较高的正负预测，并分别计算交叉熵损失，与回归的 L1 范数加权组合，得到最后的损失函数。

VoxelNet 侧重于将点云数据直接作为输入，得到检测结果，从单一的全局特征到与局部特征结合，并且逐渐从点云数据的特性来减少计算量，提高效率。相信随着计算能力的提升和研究者的不断探索，会提出更高效、更准确的方法。

5.2.3 基于视觉和激光雷达融合的障碍物检测

总体来讲，摄像头方案成本低，可以识别不同的物体，在物体高度与宽度测量、车道线识别、行人识别准确度等方面有优势，是实现车道偏离预警、交通标志识别等功能不可缺少的传感器，但作用距离和测距精度不如毫米波雷达，并且容易受光照、天气等因素的影响。毫米波雷达受光照和天气因素影响较小，测距精度高，但难以识别车道线、交通标志等元素。另外，毫米波雷达通过多普勒偏移的原理能够实现更高精度的目标速度探测，同时通过视觉可以获得充分的语义信息，而激光雷达则可以获得准确的位置信息，所以融合两种方法可以得到更好的检测效果。下面介绍几种融合方法。

1. 空间融合

建立精确的雷达坐标系、三维世界坐标系、摄像机坐标系、图像坐标系和像素坐标系之间的坐标转换关系，是实现多传感器数据的空间融合的关键。雷达与视觉传感器空间融合就是将不同传感器坐标系的测量值转换到同一个坐标系中。由于前向视觉系统以视觉为主，只需将雷达坐标系下的测量点通过坐标系转换到摄像机对应的像素坐标系下即可实现多传感器的空间同步。

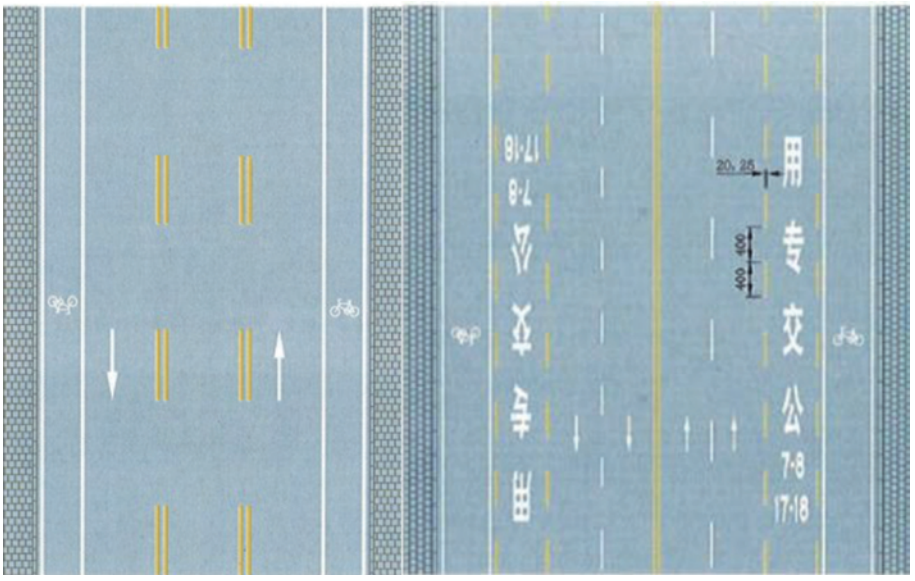
根据以上转换关系，可以得到雷达坐标系和摄像机像素坐标系之间的转换关系。由此，即可完成空间上雷达检测目标匹配至视觉图像，并在此基础上，将雷达检测对应目标的运动状态信息输出。

2. 时间融合

雷达和视觉信息除在空间上需要进行融合外，还需要传感器在时间上同步采集数据，实现时间的融合。根据毫米波雷达功能工作手册，其采样周期为 50ms，即采样帧速率为 20 帧/秒，而摄像机采样帧速率为 25 帧/秒。为了保证数据的可靠性，以摄像机采样速率为基准，摄像机每采样一帧图像，选取毫米波雷达上一帧缓存的数据，即完成共同采样一帧雷达与视觉融合的数据，从而保证了毫米波雷达数据和摄像机数据时间上的同步。

5.3 车道线检测

车道线是用来管制和引导交通的一种标线，由标化于路面上的线条、箭头、文字、标记和轮廓标识等组成。根据道路交通标志和标线国家标准(GB 5768—1999)规定，我国的道路交通标线分为指示标线、禁止标线和警告标线。道路车道线示意图如图 5-11 所示。



■图 5-11　道路车道线示意图

车道线检测是智能车辆辅助驾驶系统中必不可少的环节，快速、准确地检测车道线在协助智能车辆路径规划和偏移预警等方面尤为重要。从 20 世纪 60 年代起，车道线检测引起了广大厂商和学者的注意。目前较为常见的车道线检测方案主要是基于传统计算机视觉的检测，近年来逐渐出现了基于深度学习的道路特征预测来替代传统方法，同时随着智能交通的逐步发展，基于雷达等高精设备的车道线检测也被提出。本章将对上述三种车道线检测方法进行原理介绍和对比。

5.3.1　基于传统计算机视觉的车道线检测

传统计算机视觉的车道线检测主要依赖于高度定义化的手工特征提取和启发式的方法。国内外广泛使用的检测方法主要分为基于道路特征和道路模型两种方法。基于道路特征的检测方法主要利用车道线与道路之间的物理结构差异对图像进行后续的分割和处理，突出道路特征，实现车道线检测；基于道路模型的检测方法主要利用不同的道路图像模型（直线、抛物线、复合型），对模型中的参数进行估计与确定，最终与车道线进行拟合。本节将主要对两种方法进行介绍与讨论。

1. 基于道路特征的检测方法

基于道路特征的检测方法根据提取特征的不同，可以进一步分为基于颜色特征、纹理特征和多特征融合的检测方法。

1）基于颜色特征的检测方法

（1）基于灰度特征的检测方法。

基于灰度特征的检测方法主要通过提取图像的灰度特征来检测道路边界和道路标识。可以通过直接采集灰度图进行处理，也可以通过图像转换将原始图像转为灰度图。在车道图像中，路面与车道线交汇处的灰度值变化较剧烈，可以利用边缘增强算子突出图像的局部

边缘，定义像素的边缘强度，通过设置阈值的方法提取边缘点。常用的算子有 Sobel 算子、Prewitt 算子、Log 算子和 Canny 算子，提取效果如图 5-12 所示。

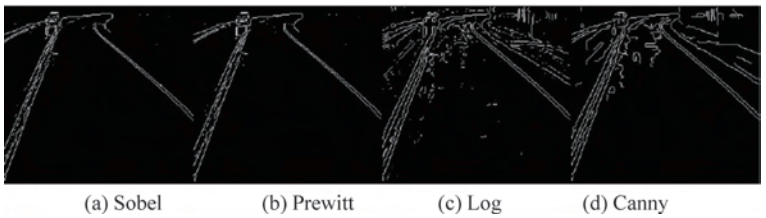


图 5-12　几种常用算子边缘提取效果示意图

这种特征提取进行检测的方法结构简单，应用广泛，对于路面平整、车道线清晰的结构化道路尤为适用。但当光照强烈、有大量异物遮挡、道路结构复杂、车道线较为模糊时，检测效果会受到较大的影响。

（2）基于彩色特征的检测方法。

基于彩色特征的检测方法主要通过提取图像的彩色特征来检测道路边界和道路标识，主要涉及颜色空间的选择和分割策略选取两方面。

颜色空间由一组数值描述图像信息的抽象模型，通常为三个数字，常用的颜色空间主要有 RGB 空间、HSI 空间和 CIE Lab 空间等。以 RGB 空间为例，通过红、绿、蓝三原色来描述图像，R、G、B 分别代表红、绿、蓝的亮度值，范围为 0～1。因此任意颜色值都可以用 R、G、B 三种颜色的分量值表示，但是 RGB 颜色空间人眼并不能直观感受到。HSI 空间用色调 H、饱和度 S、强度 I 来描述图像，将色调用角度值 0°～360°来表示。而 CIE Lab 是描述人眼可见颜色的最完备的模型，L 代表从黑到白的亮度，取值为 0～100，a 代表从绿到红的颜色区间，b 代表从蓝到黄的颜色区间，取值均为－120～120。HSI 空间和 CIE Lab 空间均可以由 RGB 彩色模型通过转换得到。图 5-13 为三种颜色空间示意图。

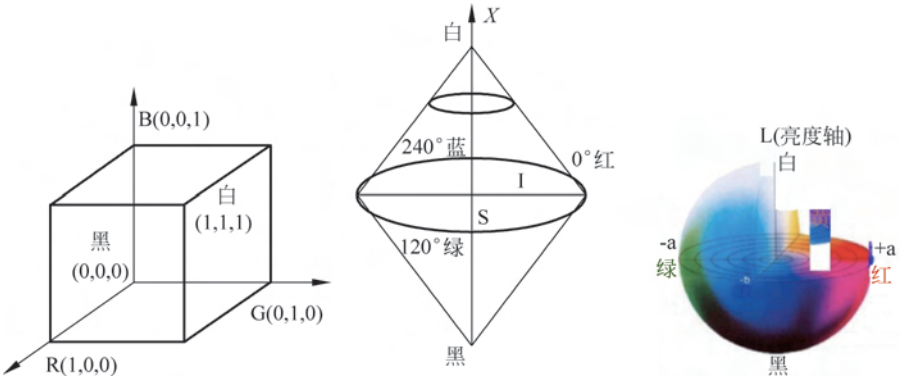


图 5-13　RGB、HSI 和 CIE Lab 三种颜色空间坐标表示示意图

在不同的颜色空间中，车道线和道路有各自的特性，通过分析彩色信息的空间分布，可以利用分割策略对车道线进行检测。通常用于车道线检测的分割策略为阈值分割和色彩聚类两种方法。

由于色彩信息对于图像或图像区域的大小、方向等特征变化不敏感，其对于局部特征，

利用彩色信息不能有效地进行捕捉，所以仅利用彩色特征的方法往往会将大量不必要的图像检测出来。

2）基于纹理特征的检测方法

基于纹理特征的检测方法主要通过对包含多个像素点的区域中的纹理强度和纹理方向进行计算，从而对车道线进行检测。这种方法具有较强的抗噪能力。但当光照强度改变、图像分辨率改变时，计算结果会有偏差。同时二维图像中提取的纹理特征与三维物体实际的纹理会有一定的差别，一定程度上影响了检测的准确度。

3）基于多特征融合的检测方法

针对单一道路特征提取的检测方法存在的缺陷，基于多特征融合的检测方法通过灵活运用多种道路特征来进行车道线检测，提高检测效果。随着图像处理技术的不断发展，越来越多的多特征融合的检测方法被提出。

2. 基于道路模型的检测方法

道路的几何模型大体分为两种：直线和曲线。直线模型计算简单，是最常用的道路模型，而曲线模型由于较为复杂，所以根据不同的情况有多种多样的模型，不同模型的计算复杂度也存在差异。

1）直线模型

直线模型主要建立在车道线为直线的假设基础上，直线模型的数学表达式如下：

$$u = k(v - h) + b$$

其中 u、v 分别代表道路图像的横纵坐标，k 代表斜率，b 为截距，h 代表道路消失线在途中的纵坐标。得到了道路消失线的水平位置后，只要得到 k 和 b 就可以确定车道线在图像中的位置。在车辆行驶速度不高，并且道路弯曲曲率不大的情况下，可以有较好的识别和导航效果。

2）双曲线道路模型

直线模型虽然实时性较好，但对曲线道路的识别精度较差。

针对一些车道线检测算法识别率不高、弯道检测不准确的问题，基于双曲线模型的车道线检测算法首先运用 Canny 算子对道路边缘进行检测；采用 Hough 变换提取道路边界点，并使用扩展的 Kalman 滤波进行预测跟踪来减小道路扫描范围；最后通过左右车道边界参数与双曲线模型参数进行匹配，利用最小二乘法来求解模型参数，完成车道边界重建。实验结果表明，这种双曲线模型的识别准确率能够达到 93.4%，并且每帧图像的处理速度为 87.4ms，在车道线模糊、对比度较低的情况下也能快速准确地识别出车道线。

5.3.2　基于深度学习的车道线检测

传统的车道线检测方法需要人工对道路场景进行特征提取和模型建立，而车道线种类繁多，道路结构复杂，传统方法工作量大且健壮性差。随着深度学习的兴起，CNN 将视觉理解推向了一个新的高度。把车道线检测看作分割问题或分类问题，利用神经网络去代替传统视觉中手动调节滤波算子的方式逐渐被提出。本节主要介绍近两年来基于深度学习的车道线检测的方法。

1. LaneNet + H-Net 车道线检测

卷积神经网络(CNN)中产生的二值化车道线分割图需要进一步分离到不同的车道线实例中。受到语义分割和实例分割中对每个像素点进行预测的启发,LaneNet 将车道线检测问题转为实例分割问题,即每个车道线形成独立的实例,但都属于车道线这一类别。H-Net 由卷积层和全连接层组成,利用转换矩阵 $\boldsymbol{H}$ 对同一车道线的像素点进行回归。

如图 5-14 所示,对于一张输入图片,LaneNet 负责输出实例分割结果,每条车道线一个标识 ID,H-Net 输出一个转换矩阵,对车道线像素点进行修正,并对修正后的结果拟合出一个三阶多项式作为预测的车道线。

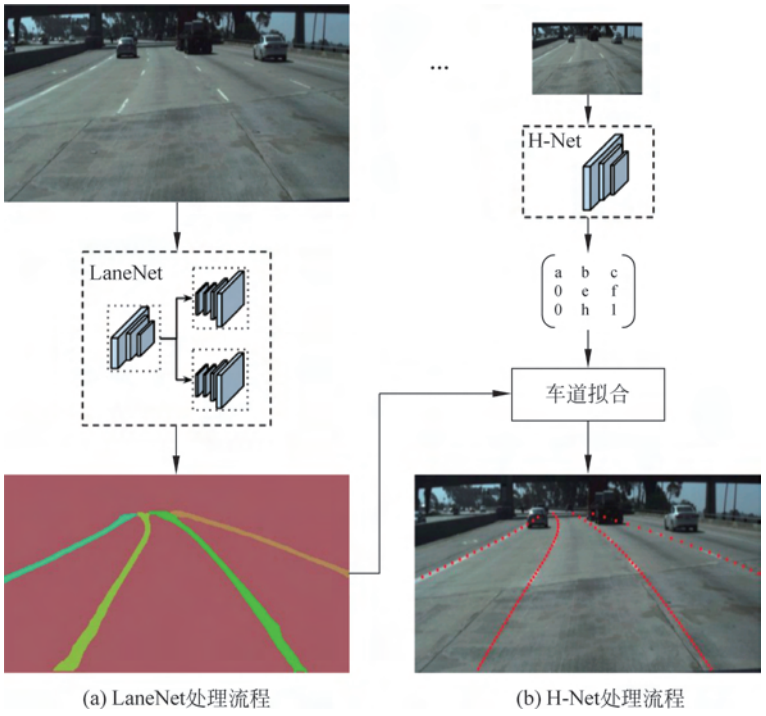
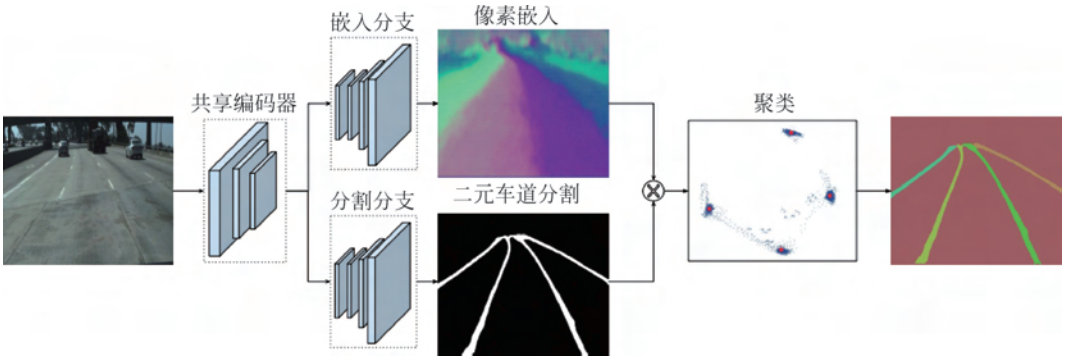


图 5-14 LaneNet + H-Net 车道线检测。图片来源于 Neven D, De Brabandere B, Georgoulis S, et al.,2018

LaneNet 将实例分割拆分为语义分割和聚类两部分,如图 5-15 所示。编码器分为 Embedding 和 Segmentation 两个分支,Embedding 负责对像素进行嵌入表示,训练得到嵌入向量进行聚类; Segmentation 对输入的图像进行语义分割,并对像素点进行二分类,判断属于车道线还是背景。最后将两个分支的结果结合得到最终车道线检测的结果。

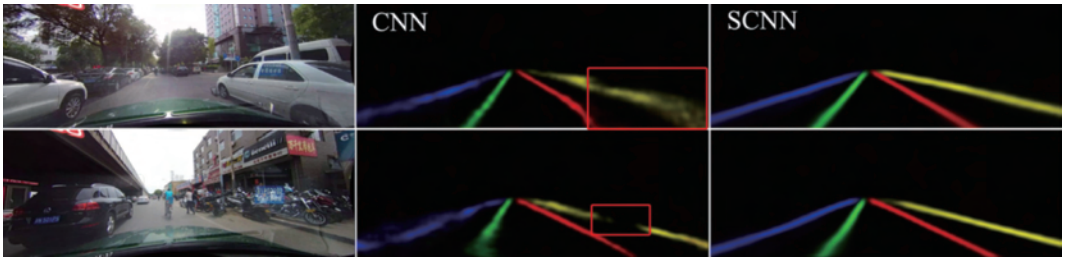
2. SCNN 车道线检测

虽然 CNN 具有强大的特征提取能力,但由卷积块堆叠的 CNN 架构没有足够充分的探索图像行和列上的空间关系的能力。而这些关系对于学习强先验形状的对象尤为重要。如图 5-16 所示,CNN 对于经常被遮挡的车道线识别效果并不好。针对这一问题,一个新的网络 Spatial CNN(简称 SCNN)将传统卷积层接层(layer-by-layer)的连接形式转为特征图中



■图 5-15 LaneNet 网络结构。图片来源于 Neven D，De Brabandere B，Georgoulis S，et al.，2018

片连片卷积(slice-by-slice)的形式，使图像中的像素行和列之间可以传递信息。SCNN 对于长距离连续形状的目标、大型目标以及有着极强空间关系但是外观线索不明显的目标，例如车道线、电线杆，具有很好的检测效果。如图 5-16 所示，SCNN 对遮挡的车道线的检测的结果明显好于传统 CNN 的方法。



■图 5-16 CNN 与 SCNN 对于车道线检测效果。图片来源于 Pan X，Shi J，Luo P，et al.，2018

在此之前，有很多尝试在深度网络中使用空间信息的工作。例如，有使用循环神经网络(RNN)按每行或每列传递信息，但每个像素点只能接收来自同一行或同一列的信息；有使用长短期记忆网络(LSTM)的变体来探索语义分割中的上下文信息，但计算量较大；也有使用 CNN 和图模型，如马尔可夫随机场(MRF)和条件随机场(CRF)结合，通过大卷积核来传递信息。

与上述几种方法相比，SCNN 在信息传递过程中计算效率比 MRF 和 CRF 高，同时由于使用残差进行信息传递，使训练更容易进行并适用于多种神经网络。

传统 CNN 与图模型结合时，每个像素点接收来自整个图像的其他像素的信息，这在计算上的代价是非常昂贵的，不利于应用与车道线的检测。并且 MRF 大卷积核的权重在学习上也比较困难。如图 5-17 所示，SCNN 结构中，D、U、R、L 在结构上是类似的，分别代表向下、向上、向右、向左。

以 SCNN_D 为例，对于 $C\times H\times W$ 的三维张量，C、H、W 分别代表通道数、高度和宽度，先将其切分为 H 片，然后将第一片送入 $C\times w$ 的卷积层中(w 代表卷积核的大小)。与传统 CNN 将输出作为下一片传递到下一层不同的是，SCNN 将第一片的输出加入到下一片中作为输入，重复卷积，直到处理完最后一片。SCNN_U、SCNN_R、SCNN_L 在处理上采用的方法与上述方法基本一致。

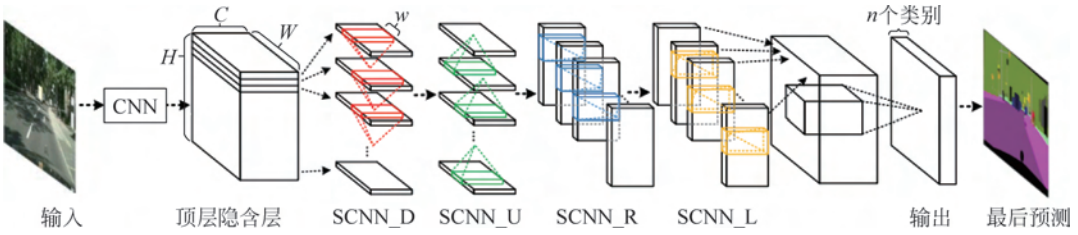


图 5-17 SCNN 结构图。图片来源于 Pan X，Shi J，Luo P，et al.，2018

SCNN 在信息传递上与 MRF/CRF 的不同之处主要在于信息传递的方向，如图 5-18(a)所示，MRF/CRF 在信息传递方向上较为无序，每个像素点需要接收大量的信息，计算量大，存在大量冗余信息；而 SCNN 信息的传递是顺序的，如图 5-18(b)所示，对于行列数较高的图片，SCNN 可以减少大量的计算。图 5-19 展示了 SCNN 与其他方法在车道线检测中的对比。

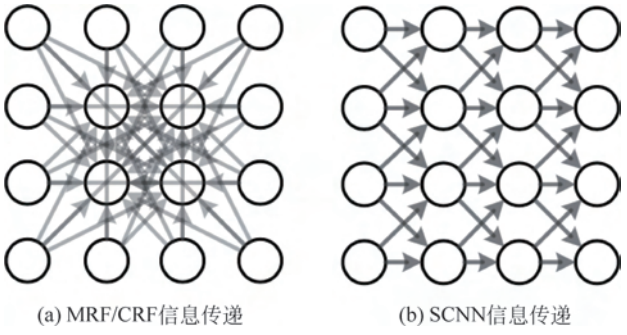


图 5-18 MRF/CRF 信息传递和 SCNN 信息传递。图片来源于 Pan X，Shi J，Luo P，et al.，2018

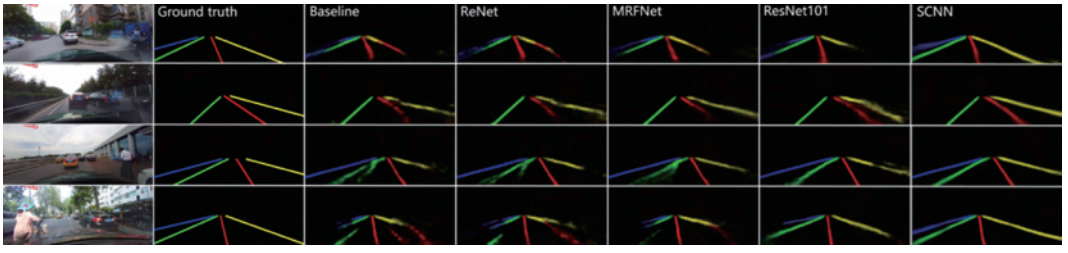


图 5-19 SCNN 与其他方法在车道线检测中的表现。图片来源于 Pan X，Shi J，Luo P，et al.，2018

5.3.3 基于激光雷达的车道线检测

基于传统视觉的方法存在诸多缺陷：对光照敏感、依赖于完整并且较为统一的车道线标识、有效采样点不足以及车道线被水覆盖时视觉系统会失效等。近年来，越来越多的研究者将目光投向了用激光雷达进行车道线检测。激光雷达的有效距离比传统视觉高，有效采样点多，并且可以穿透水面，基本上解决了传统视觉中的大部分问题，但基于激光雷达的检测方法存在“硬伤”：成本较高。本节主要介绍一种基于激光雷达的车道线检测方法，即基

于反射强度信息的方法。

该方法主要基于激光雷达反射强度信息形成的灰度图，或者根据强度信息与高程信息配合，过滤出无效信息，然后对车道线进行拟合。不同物体的回波强度见表 5-1。

表 5-1 不同物体回波强度

介　　质	回波强度/dBz	可能的物体分类
沥青、混凝土	5～8	道路、房屋等
特性涂层	12～30	车道线
植被、金属	45～150	树木、车辆等

在激光雷达获取的点云中，通过反射强度值，可以区分出道路和车道线。在激光雷达获取的道路环境的三维点云中，检测每一个激光层采集到的可行驶区域的回波强度是否发生变化，如果发生变化，将变化点提取并进行标记。

与此同时，通过对点云数据中有高程数据的点进行滤波，一定程度上可以确定出可行驶区域，同时剔除一些和车道线的回波强度接近的物体。通过对提取的车道线点云进行聚类和去噪，再利用最小二乘法进行拟合，最终提取出车道线。

5.4 红绿灯检测

红绿灯检测是无人驾驶中一个关键的问题，红绿灯检测就是获取红绿灯在图像中的坐标以及它的类别。无人驾驶汽车根据检测的结果采取不同的措施：如果检测到红灯，则在路口等待；如果检测到绿灯，则通过路口。因此，能否准确识别红绿灯的状态，决定着无人驾驶汽车的安全。

过去对于红绿灯的检测，大多都是利用颜色形状等低级特征去做，例如在颜色上使用一个简单的阈值进行背景抑制，或者是根据颜色特征进行候选框的提取，再对候选框进行分类，这类方法准确率远远达不到要求。最近几年计算机视觉大量使用深度学习的技术，目标检测的准确率和速度有了很大的提升。

现在大多目标检测的方法都是基于 Faster RCNN、YOLO 和 SSD，但是它们在小目标检测上的效果都不理想。红绿灯这种小目标在图片中所占据的像素较少，对于标准的卷积神经网络（VGG、ResNet、DenseNet 等）来说，输出的特征一般都会是图片大小的 1/32，对于小目标来说，细节丢失较严重，这就增加了小目标检测的难度。如果删除特征提取网络的一些层数或者部分下采样层，就会缩小感受野，衰弱特征的语义信息，反而更影响检测效果。

对于 Faster RCNN 和 SSD 这种基于候选窗口的检测算法来说，为了完美覆盖小目标，需要候选窗口数量较多，这不仅拖慢了检测速度，而且增加了前背景框的分类难度。

提升小目标检测效果的最有效方法是扩大图像大小，而扩大图像大小随之带来的就是计算量的增加，感受野的不足影响大目标的检测效果。目前针对小目标检测算法的改进，主要从提取特征网络入手，让提取的特征更加适合小目标检测，大致有如下几种方法。

（1）图像金字塔：较早提出对训练图片上采样出多尺度的图像金字塔。通过上采样能

够加强小目标的细粒度特征，在理论上能够优化小目标检测的定位和识别效果。但基于图像金字塔训练卷积神经网络模型对计算机计算能力和内存都有非常高的要求。计算机硬件发展至今也难以胜任。故该方法在实际应用中极少。

（2）逐层预测：对于卷积神经网络的每层特征图输出进行一次预测，最后综合考量得出结果。这种方法会利用浅层特征去做预测，而浅层特征没有充分的语义信息，也没有较大的感受野，所以效果对于前面的特征层来说，并不会因为特征图的变大而变好。同样，该方法也需要极高的硬件性能。

（3）特征金字塔：参考多尺度特征图的特征信息，同时兼顾了较强的语义特征和位置特征。较大的特征图负责较小的目标的检测，较小的特征图负责较大的目标检测。该方法的优势在于，多尺度特征图是卷积神经网络中固有的过渡模块，虽然在提取特征时候增加了部分层，但是尺度的特征通道也会减少，所以堆叠多尺度特征图对于算法复杂度的增加微乎其微。

（4）空洞卷积：利用空洞卷积代替传统的卷积，在提升感受野和不增加额外参数的同时，不减少特征图的大小，保留更多的细节信息。

（5）RNN 思想：参考了 RNN 算法中的门限机制、长短期记忆等，同时记录多层次的特征信息，但是 RNN 固有的缺陷是训练速度较慢。

5.4.1 基于传统视觉方法的红绿灯检测

1. 基于颜色和边缘信息

在传统的视觉任务中，一般都会使用 HSV 或者 RGB 颜色空间，但是在红绿灯检测任务中，往往颜色和边缘信息更有效。

该方法的步骤是：①获取图像帧；②图像预处理，将 RGB 转换为 CIE Lab 颜色域、增加红绿颜色差距、填充空洞；③候选区域检测，径向对称检测、最大最小定位；④候选区域验证：时空持续性验证。大致过程如图 5-20 所示。

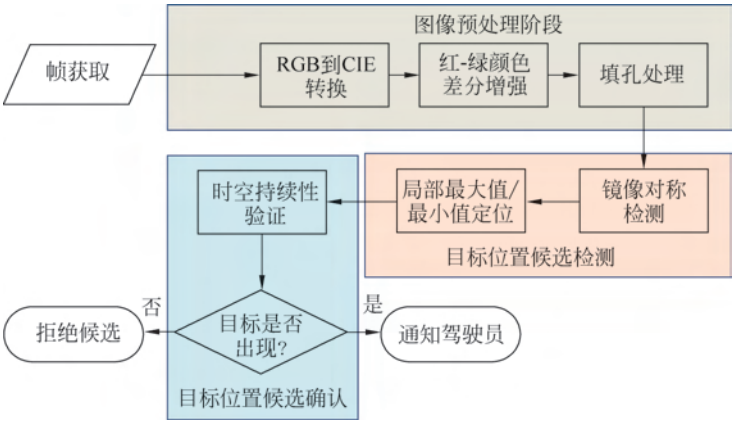


图 5-20 算法流程图

如图 5-21 所示，该方法利用简单的颜色和边缘信息取得了不错的效果。该方法虽然在速度上非常快，但是该算法健壮性不足，在一些场景中容易发生误检。

图 5-21　算法效果图

2. 基于背景抑制

基于背景抑制的方法就是通过处理图像浅层特征来区分前景与背景，从而实现背景抑制和红绿灯候选区域获取。一个好的二阶段检测算法必须要能够提取高质量的候选框，只要候选框够准，红绿灯检测的召回率就高。该方法包含两个部分：候选区域提取模块以及识别模块。具体结构如图 5-22 所示。在候选区域提取模块中，该方法使用自适应背景抑制去突出前景从而获取候选区域。在识别模块中，每个候选区域的特征输入到识别网络中，获得候选框的类别(红灯、绿灯以及背景)。算法结构如图 5-22 所示。

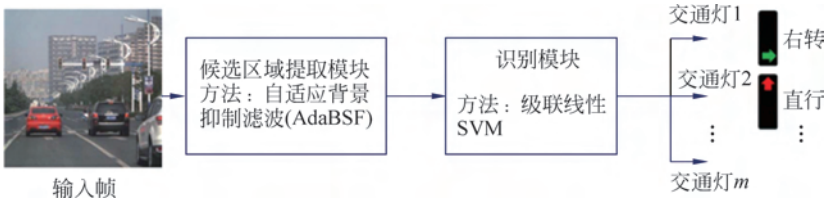


图 5-22　算法结构图

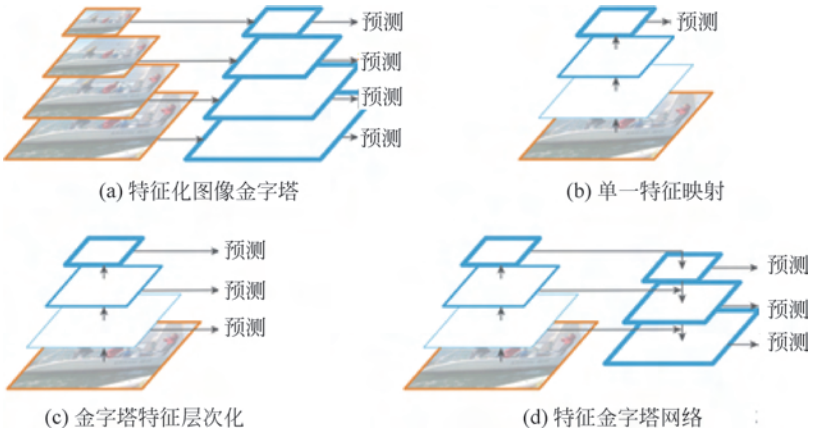
与基于颜色和边缘信息的方法相比，而该方法是自适应的方法，所以能够适应多种场景，并且具有很高的召回率。

5.4.2　基于深度学习的红绿灯检测

1. 特征金字塔网络(FPN)红绿灯检测

在传统的图像处理方法中，金字塔是比较常用的一种手段，如 SIFT(尺度不变特性变换)算法基于金字塔做了多层的特征采集。对于深度网络来讲，其原生的卷积网络特征决定了天然的金字塔结构。深度网络在目标检测领域的应用，如早期的 Fast RCNN、Faster RCNN，都是在最后一层卷积层进行检测，后续针对的改进包括 ION、HyperNet、MSCNN 等都结合了多尺度的特征。

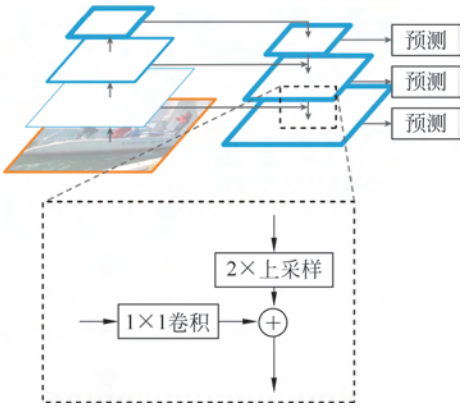
现在处理不同尺度特征的方法如图 5-23 所示。图 5-23(a)是传统方法，通过对图像进行降采样处理，提取每层图像的特征，然后在每层进行预测。图 5-23(b)是借助卷积网络，通过单特征图进行预测，典型的应用包括 Faster RCNN、YOLO 等。图 5-23(c)是对不同尺度的特征图分别进行预测。图 5-28(d)是特征金字塔方法，在多尺度特征图的基础上，结合右侧的上采样进行不同尺度的整合，每层独立预测，通过本层信息和原始特征层信息进行结合。



■图 5-23 多种特征处理方式。图片来源于 Lin,et al.,2017

Feature Pyramid Networks (FPN) 是比较早提出利用多尺度特征和从上到下结构做目标检测的网络结构之一,整个网络是基于 Faster RCNN 检测算法构建的。

在原始的 Faster RCNN 中,只用网络高层特征去做检测,虽然语义信息比较丰富,但是经过下采样等操作,特征丢失太多细节信息,而对于小目标检测这些信息往往是比较重要的。所以,该方法想要将语义信息充分的高层特征映射回分辨率较大、细节信息充分的底层特征,将二者以合适的方式融合来提升小目标检测的效果,融合方式如图 5-24 所示,将高层特征利用上采样的方式转化成和低层特征的相同尺寸,同时两者通道数相同,再将低层和高层特征进行元素级相加。



■图 5-24 特征上采样结构图。图片来源于 Lin,et al.,2017

该方法可以套用在各种目标检测算法上,已经成为现在目标检测算法中的一个标准组件,一定程度上平衡了不同大小目标的检测,有着不错的精度,速度上也快于图像级特征金字塔。

2. 特征融合 SSD 红绿灯检测

上述 FPN 是在二阶段算法基础上实现的,特征融合 SSD 则是在一阶段检测算法 SSD 基础上对小目标检测做的一些改进,该方式使用多尺度特征融合将上下文信息引入到 SSD 中帮助检测红绿灯这种小目标。上下文信息对于小目标检测的重要性不言而喻,红绿灯这种小目标和背景之间的尺寸差异大,如果使用较小的感受野去关注物体本身的特征,则很难提取到背景中包含的全局语义信息。如果使用较大感受野去关注背景信息,那么小目标本身的特征就会被丢失,所以使用多尺度特征融合可以有效解决这一问题,浅层网络得到精细特征,高层网络通过大的感受野得到上下文信息,两者相结合,从而改善小目标检测,同时也不会降低大目标检测的效果。

该方法基于原始的 SSD 结构,整体结构如图 5-25 所示。

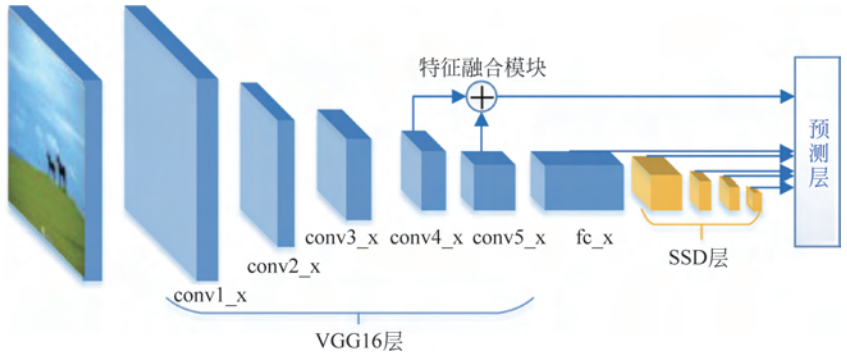


图 5-25 特征融合 SSD 网络结构图。图片来源于 Li,et al.,2017

SSD 是基于 VGG16 的基础网络,仅仅是替换掉 VGG16 中的一些层以及增加一些网络层。SSD 不使用最后一个特征映射去做预测,而是使用卷积层中的多层中的金字塔特征层次结构来预测具有不同规模的目标。也就是说,使用浅层来预测较小的对象,同时使用深层来预测较大的对象,这样可以减少整个模型的预测负担。然而,较浅的层往往缺少语义信息,这是较小的对象检测的重要补充,因此,将深层得到的语义信息传递回浅层可以提高小目标的检测性能,同时也不会增加太多的计算量。

通过将浅层与深层的特征融合,可以为小目标检测提供丰富的上下文信息。在检测中,深层特征往往由于太大的感受野通常会引入大的无用的背景噪声,浅层特征则因为网络不够深、没有充足的语义信息,所以该方法利用浅层与深层融合之后的特征对小目标进行预测,同时为了不降低速度,对于较大的红绿灯目标,我们不使用特征融合模块,而是直接使用较深的高级特征。在小的红绿灯检测上为了选择合适的特征融合层,利用反卷积针对不同尺寸的目标生成不同尺寸的特征。

该网络添加了级联模块,为了使 conv5_3 层的特征图与 conv4_3 层的特征图大小相同,在 conv5_3 层后面跟着一个反卷积层,该反卷积层通过双线性上采样放大特征尺寸。在 conv4_3 层和 conv5_3 层之后,使用两个 3×3 的卷积层,以学习更好地融合特征,最终的融合特征图由 1×1 卷积层生成,用于降维和特征重组。另一种融合两层特征图的方法是使用元素级相加模块。该模块除了融合类型之外,都与级联模块相同。事实上,由于这种方式能够自适应地从 conv4_3 和 conv5_3 学习特征,所以可以获得更好的融合效果。

特征融合 SSD 是一阶段的方法,在 SSD 的基础上进行了针对小目标检测的优化,使卷积特征更适合在无人驾驶中的红绿灯检测,与二阶段算法相比,在满足实时性的同时,也有不错的准确率。

5.4.3 高精地图结合

上述检测算法只是通过图像获取红绿灯在图像中的位置,而获取红绿灯世界坐标则需要结合高精地图。高精地图是指高精度、精细化定义的地图,其精度需要达到分米级才能够区分各个车道。如今随着定位技术的发展,高精度的定位已经成为可能。而精细化定义则是需要格式化存储交通场景中的各种交通要素,包括传统地图的道路网数据、车道网络数据、车道线和交通标志等数据。

在利用计算机视觉进行红绿灯检测时，必须在整幅图像中搜索，因为计算机是无法预知红绿灯出现在图像中的具体位置的。但是如果有了高精度地图信息，机器就可以通过高精度定位和高精度地图得到 ROI。根据定位和地图的数据，无人驾驶汽车可以知道前方、两侧是否有交通标志牌，及红绿灯的位置，这样就可以大幅度降低算法的复杂度，减少系统的计算负荷，进而提升系统性能。

高精地图与红绿灯检测的具体结合模式如图 5-26 所示，首先通过使用检测算法，确定红绿灯在图像中的位置以及它的类别，然后将红绿灯与高精地图上记录的红绿灯进行比对(map matching)，比对之后无人驾驶汽车就可以得到红绿灯的世界坐标，确定红绿灯所对应的道路，从而帮助无人驾驶系统做出正确的决策。当无人驾驶汽车因为遮挡或者算法等原因无法检测到红绿灯时，高精地图可以告知系统红绿灯的信息，从而确保行车安全。

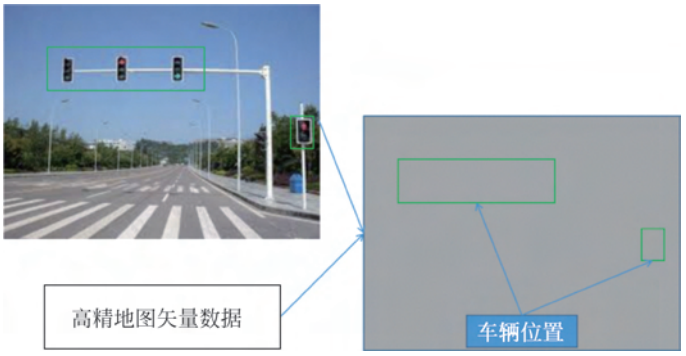


图 5-26　高精地图与红绿灯检测的具体结合模式

注：右侧图为俯视图，表示红绿灯与车辆的相对位置。

5.5 场景流

5.5.1 概述

场景流(scene flow)可以理解成空间中场景的三维运动场，即空间中每一点的位置信息和其相对于摄像头的移动。具体地，场景流估计的一种方式是光流估计和深度估计的结合。本节将重点介绍深度估计与光流估计的相关知识。

光流是一种二维运动场，是空间中每一点沿摄像头平面的运动状态；深度信息表达的是空间中每一点到摄像头的距离，其变化量则是物体沿垂直摄像头方向的变化。光流和深度变化可以认为是对三维运动场的一种分解。综上所述，对光流和深度估计的实现可以使我们对空间中任意点的三维运动状态都了如指掌，这对于驾驶决策等是很有意义的。同时，场景流的潜在应用很多。它可以补充和改进最先进的视觉测距和 SLAM 算法，这些算法通常假定在刚性或准刚性环境中工作。另外，它可以用于人-机器人或人-机交互，以及虚拟和增强现实等。但是，需要注意的是，场景流只关注深度变化量，不关注深度的绝对值。

5.5.2 深度估计

深度估计是指获取图像上每一点距离测量平面的深度信息，在无人驾驶中可用于障碍物的识别与定位，由于它能够获取障碍物与测量点的距离信息，可用于三维重构、即时定位与地图构建(SLAM)等，在无人驾驶中有广泛的应用。

1. 基于激光雷达的深度估计

一种比较直接的方法是通过激光雷达采集点云数据，这些点云数据直接代表深度信息。这种方法采集到的深度信息直接、可靠，但是也存在某些问题。由于其稀疏性需要使用插值等方式使其稠密化，而且采集范围有限，仅限于某些固定的角度，车载时有时速范围限制，还有价格昂贵等。如图 5-27 所示，这是一张由激光雷达采集得到的深度图。可以看到，这个深度图是由多条线构成的。这是因为采集所使用的激光雷达是多线的(例如 32 线、64 线)，这样导致扫描结果以线的形式呈现，而不是连续致密的结果。而且可以看出，这幅图上面一部分是没有数据的，这是因为激光雷达的扫描范围限制。因此，我们也希望使用光学图像来测量深度，也就是基于光学图像的深度估计，进而能够实现多传感器协同工作，提高系统的健壮性。

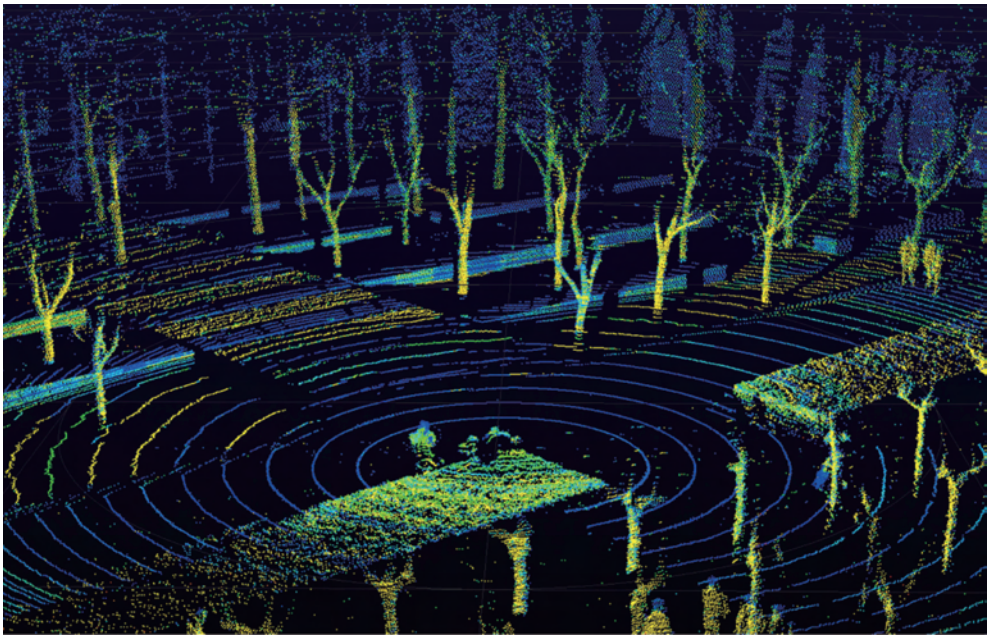


图 5-27 激光雷达采集的深度图

2. 基于光学图像的深度估计

基于光学图像的深度估计可分为双目深度估计和单目深度估计。

(1) 双目深度估计：此方法和人类感知环境深度的方式较为类似，通过两个具有一定相对位置的摄像头(可类比人的两只眼睛)来采集光学图像，然后通过这两张光学图像来确定深度信息，这可以归纳为一个立体匹配的问题，也就是找到左图(左眼)和右图(右眼)中对应点，利用对应点的视差(空间中同一点在左图和右图中成像位置的变化量)来确定该点的

深度。如图 5-28 所示，要确定空间中一点 P 到两摄像头主点(O_R，O_T)连线的距离 Z。由图中几何关系可知，由于左摄像头和右摄像头的位置是不同的，导致 P 点分别在左摄像头焦平面和右摄像头焦平面上成像，在 p 和 p' 点，定义视差为 $d=x_R-x_T$，它为点 P 在左图和右图中成像位置的差异。焦距为 f，两摄像头主点距离为 $B=O_T-O_R$。图中存在相似三角形 $P_{pp'}$ 和 PO_RO_T，即有 $\frac{Z-f}{Z}=\frac{B-d}{B}$，进而得到

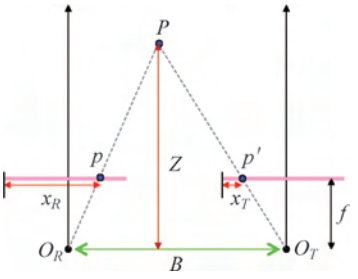


图 5-28 双目深度估计原理图

$$\frac{f}{Z}=\frac{d}{B} \quad Z=\frac{f\times B}{d}$$

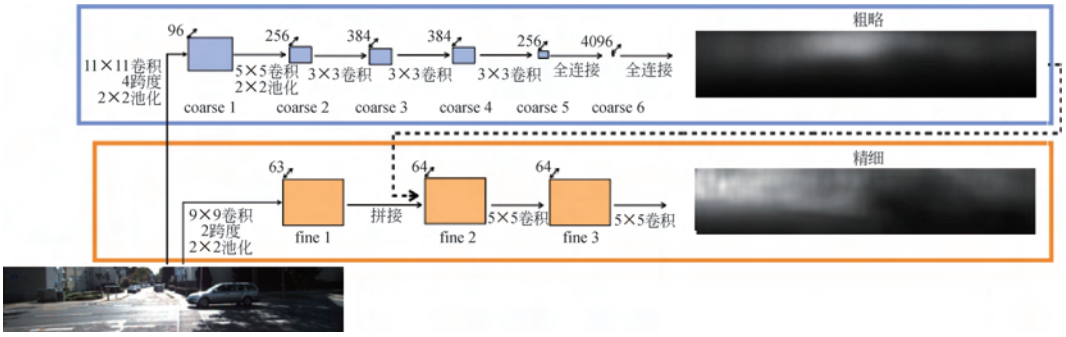
由上式可知，在摄像头焦距(f)和两摄像头排放位置相对固定(主点距离 B 为常数)时，空间中一点的成像视差和其到摄像头平面的距离成反比。这一个关系非常重要，把深度转换成了视差，所以只需要确定左图和右图上对应的点，就可以根据这些对应点的视差确定深度。

(2) 单目深度估计：其装置简单，便于安装，应用前景也相对较大，是目前一个较为火热的研究方向，其包括基于图像内容理解、基于聚焦、基于散焦、基于明暗变化等方式。严格来讲，单目深度估计的信息量是不够的，因为图像上的一点对应于空间中的一条射线，理论上无法直接通过一张图片定位空间中点的位置，就好像人闭上一只眼睛来观察周围的环境，会明显地感觉到缺少了很多立体感。但是，还是可以通过仅一只眼睛工作和生活等，而不至于撞到墙上，这是由于我们的经验，我们在长期的生活中已经建立了对各种事物的认知，可以从物体的观测大小和其周围环境推断出距离。例如，我们观察到了公路上的一辆汽车，经验告诉我们它的大致尺寸是 1.5m×1.7m，那么我们也能轻易地推断出它在多远的位置应该有多大的观测大小，当它在我们的视线里越来越大时，那么它离我们的距离就越来越近。由于单目深度估计需要基于语义理解，而在场景发生剧烈改变时可能会发生严重的错误。例如，还是在上面的例子中，我们又观测到了一辆车，这辆车只是一个模型，它的实际大小只有 20cm×30cm。但是，我们把它当成了一辆真实的车，那么我们对它距离的推断要远远大于它的实际距离，而这是因为我们错误地认识了它。基于以上分析，单目深度估计的实现往往需要一致的环境，相同的数据分布。

早期的单目深度估计采用直接回归深度的方法，通过卷积神经网络直接回归深度信息，具体流程如图 5-29 所示。

输入一张图片，通过一个 CNN 生成一个粗略的结果，如图 5-29 中蓝色框流程所示。蓝色 CNN 中有卷积层和下采样，感受野很大，最终输出全局的整体结果。然后将这个粗略的结果与原图一起输入下面黄色的精细网络里。下面的网络中没有下采样的过程，目的是根据原图和粗略的结果提炼细节和边缘信息，进而得到精细的结果。这种方法的效果一般，而且鉴于深度信息难以获得，较难获得标注数据，其应用场景受限。

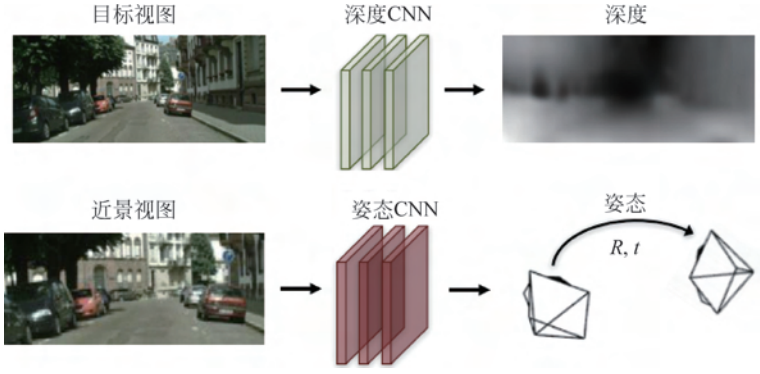
除了直接估计以外，单目深度估计往往会用到重构的方法，通过与原图空间相关或者时序相关的另一张图进行估计。这种方式模型构造较为复杂，需要利用空间关系进行重采样，但是效果相较于直接利用深度信息估计的方法好很多。最重要的是，这种方法往往不需要



■图 5-29 通过卷积神经网络直接估计深度流程

真实的深度信息，也就是不需要拍摄视频的同时使用激光雷达采集深度信息，大大降低了数据获取的成本，具有很高的实用价值。下面介绍两种基于重构的单目深度估计方法：一种是基于视频的方法，重构视频中的前后帧；另一种是给定左图重构右图的方法（只是在训练的时候需要成对的图像，在测试的时候只需要一张图，所以还是将其归于单目深度估计）。

深度和自身运动网络是一种基于视频的单目深度估计方法，可以分解为两个子问题，分别是预测出每一帧的深度图（depth）和车辆自身的运动状态（ego-motion）。这对后续的三维重构等任务很有帮助。因此，本方法的实现模型可分为两个子网络：预测深度的网络和预测摄像头自身运动的网络，如图 5-30 所示。



■图 5-30 预测深度网络和预测自身运动网络示意图。图片来源于 Zhou，et al.，2017

从前面章节讲摄像头参数的内容中可以得知，空间坐标系中一点到摄像头坐标系的映射关系为

$$\boldsymbol{Z}\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \boldsymbol{KM}\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

其中，(X,Y,Z)是空间坐标系中的一点，$(x,\ y)$是该点在摄像头坐标系中的坐标，$\boldsymbol{M}$ 是摄像头外参，$\boldsymbol{K}$ 是摄像头内参。

假设知道 t 时刻某张图的深度信息，也就是已知该图上某点坐标(x,y)，以及该点对应

的深度值 Z，则有

$$\boldsymbol{M}\begin{bmatrix}X\\Y\\Z\\1\end{bmatrix}=\boldsymbol{Z}\boldsymbol{K}^{-1}\begin{bmatrix}x\\y\\1\end{bmatrix}$$

若已知 t 时刻到 $t+1$ 时刻摄像头的运动状态(绕三个轴的旋转和沿三个轴的平移)，则此运动可表示为一转移矩阵 $\boldsymbol{T}^t=\begin{bmatrix}\boldsymbol{R}^t & \boldsymbol{T}^t\\0 & 1\end{bmatrix}$，则 $t+1$ 时刻的摄像头外参可表示为 $\boldsymbol{M}^{t+1}=\boldsymbol{M}\boldsymbol{T}^t$，则 $t+1$ 时刻的映射关系可表示为

$$\begin{aligned}\boldsymbol{Z}^{t+1}\begin{bmatrix}x^{t+1}\\y^{t+1}\\1\end{bmatrix}&=\boldsymbol{K}\boldsymbol{M}^{t+1}\begin{bmatrix}X\\Y\\Z\\1\end{bmatrix}\\&=\boldsymbol{K}\boldsymbol{T}^t\boldsymbol{M}\begin{bmatrix}X\\Y\\Z\\1\end{bmatrix}\\&=\boldsymbol{K}\boldsymbol{T}^t\boldsymbol{Z}\boldsymbol{K}^{-1}\begin{bmatrix}x\\y\\1\end{bmatrix}\end{aligned}$$

整理可得

$$\begin{bmatrix}x^{t+1}\\y^{t+1}\\1\end{bmatrix}=\frac{1}{\boldsymbol{Z}^{t+1}}\boldsymbol{K}\boldsymbol{T}^t\boldsymbol{Z}\boldsymbol{K}^{-1}\begin{bmatrix}x\\y\\1\end{bmatrix}\tag{5-1}$$

从式(5-1)可知，如果知道了 t 帧、$t+1$ 帧的深度信息，以及从 t 帧到 $t+1$ 帧中摄像头的自身运动状态，那么就可以找到 $t+1$ 帧中的某一点 (x^{t+1},y^{t+1}) 在 t 帧中的位置 (x,y)(前提是环境相对于世界坐标系是静止的)，进而可以从 t 帧的图像中采样得到 $t+1$ 帧图像的重构图。可以用这张重构图与 $t+1$ 帧真实的图像做监督，进而优化深度估计和自身运动估计的性能。

完整过程如图 5-31 所示，I_t 表示 t 时刻的输入图像，由深度 CNN 生成深度信息，如图中第一条流程所示。I_{t-1}、I_{t+1} 分别为 $t-1$ 时刻的图像和 $t+1$ 时刻的图像，分别与 I_t 经过姿态 CNN 预测出 $t-1$ 到 t、t 到 $t+1$ 的外参转移矩阵 $\boldsymbol{T}_{t-1\to t}$、$\boldsymbol{T}_{t\to t+1}$，然后，根据式(5-1)可得 I_{t-1}、I_{t+1} 与 I_t 的采样关系，进而从 I_t 中采样(采样过程如图 5-31 中红色虚线所示)，从而得到重构的两张图 I'_{t-1} 和 I'_{t+1}。然后利于原始图像 I_{t-1} 和 I_{t+1} 建立监督：$\|I_{t-1}-I'_{t-1}\|$ 和 $\|I_{t+1}-I'_{t+1}\|$。

图 5-32 是采样重构的具体过程，表示的是从 I_s 中采样，重构 I_t。已知 I_t 与 I_s 中点的映射关系：$p_t\to p_s$，那么在 p_s 处采用双线性插值的方法，取其临近的四个点的像素值，根据 p_s 点到每一个点的距离进行加权平均，得到 p_s 点处的像素值，作为重构图中 p_t 处的像素值。这种方法既能使重构的图像连续平滑，又能保证重构图可导，可以进行反向传播。

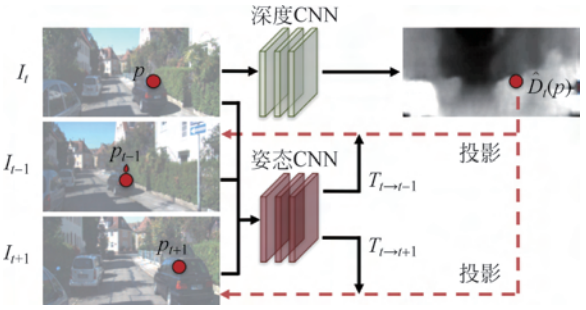


图 5-31 完整过程示意图。图片来源于 Zhou，et al.，2017

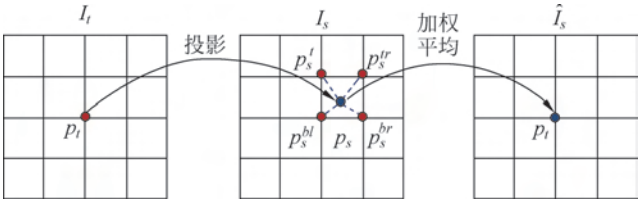


图 5-32 采样重构具体过程图。图片来源于 Zhou，et al.，2017

以上就是该方法完整的过程，该方法只需要单个摄像头所采集到的数据进行训练，而且不需要深度信息，训练数据极易获得，测试时还可逐帧测试，不需依赖视频信息。另外，该方法所预测的车辆自身运动状态对于其他任务，例如三维重构、SLAM、视觉里程计等，也是很有帮助的。

左右一致性深度估计是一种单目无监督深度估计方法。训练所使用的数据不再是连续的视频，而是成对的图像，成对的图像由两个相对位置固定的摄像头采集，保证视差与深度的固定关系为 $Z=\frac{f\times B}{d}$，其中 Z 是深度，d 是视差，f 是焦矩，B 是两个摄像头的光心距离。

虽然此方法在训练的时候用到了成对的图像，但是在测试的时候只需要其中的一张图像，所以还是将其认为是单目深度估计的一种方法。此方法也利用重构，但重构的是同一时刻不同摄像头拍摄的图像，并且不需要对深度信息进行标定。

该方法过程如图 5-33 所示，给定一组输入图片 L、R，让 L 图经过 CNN 得到预测的两个视差图，这两个视差图中 L_d、R_d 分别是左图相对于右图的视差和右图相对于左图的视差。然后，与基于激光雷达的深度估计方法中的采样方法相同，采用双线性插值方法，由左图 L 和右图相对于左图的视差图 R_d 采样重构出右图 R′，然后再由右图 R 和左图相对于右图的视差图 L_d 重构出左图 L′。最后，分别用 L 和 R 对 L′和 R′进行估计。这是整个训练的过程。

由于深度信息与视差的固定关系，在测试时只需要得到视差图就可以，不需要后面的重构过程，所以测试流程可以简化如下，对于一张给定给的左图 L，由训练好的 CNN 得到一张左视差图 L_d，然后由 $Z=\frac{f\times B}{d}$ 得到深度信息，如图 5-34 所示。

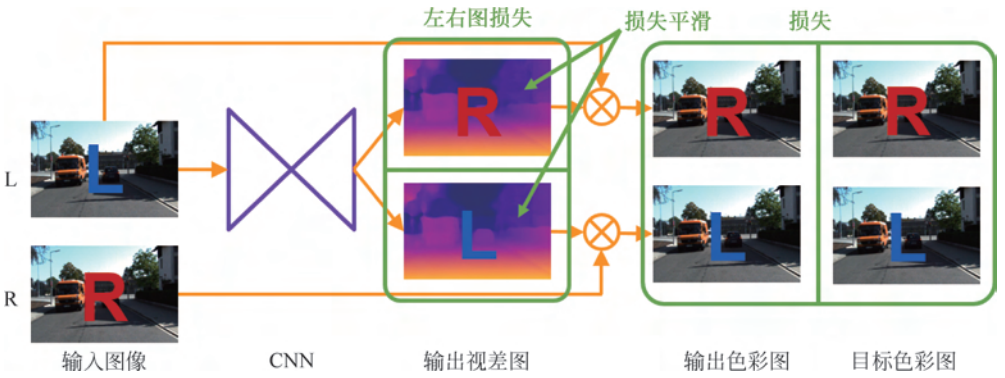


图 5-33 深度估计过程示意图

图 5-34 测试过程流程图

5.5.3 光流估计

光流指的是图像中每个像素点的二维瞬时速度场，其中的二维速度指的是物体空间中三维速度向量在成像平面上的投影。通俗地说，就是图像中的每一个像素点在图中的移动速度。光流是目前运动图像分析的一种重要方法，表达了图像变化。光流信息可以应用于动作识别、物体轨迹预测、动目标识别等，在无人驾驶领域有着重要的应用场景。

1. LK 算法

LK(Lucas-Kanade)算法是一种稀疏的光流算法，它首先对光流特性有以下几条假设：首先，运动物体的灰度值在短时间内保持不变，这也是寻找两帧之间对应点的关键所在；其次图像随时间变化较慢，也就是可以使用相邻像素点的灰度差异来表征某一点的梯度，这是光流法中极其重要的假设；最后图像的每一小邻域中光流近似一致。基于这些假设，可以进行以下分析。

假设图像上的一个像素点(x,y)，它在某一时刻的灰度为$I(x,y,t)$，用u和v分别表示该点的光流在水平和垂直方向上的分量。那么

$$u=\frac{\mathrm{d}x}{\mathrm{d}t},\quad v=\frac{\mathrm{d}y}{\mathrm{d}t}$$

经过一小段时间间隔Δt后，该点在$t+\Delta t$时的对应位置的灰度为$I(x+\Delta x,y+\Delta y,t+\Delta t)$，利用泰勒展开式可得

$$I(x+\Delta x,y+\Delta y,t+\Delta t)=I(x,y,z)+\frac{\partial I}{\partial x}\Delta x+\frac{\partial I}{\partial y}\Delta y+\frac{\partial I}{\partial t}\Delta t+\mathrm{H.O.T.}$$

其中，H. O. T. 表示更高阶的量，可忽略。再基于我们之前的第一个假设：运动物体的灰度值在短时间内保持不变，也就是 $I(x+\Delta x, y+\Delta y, t+\Delta t) \approx I(x, y, z)$，即

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0$$

当 Δt 趋于 0 的时候，可得

$$-\frac{\partial I}{\partial t} = \frac{\partial I}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial I}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} = \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v$$

$$-I_t = \begin{bmatrix} I_x & I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

其中，I_t、I_x、I_y 分别是灰度相对于时间、横坐标、纵坐标的导数。在基于前面假设的基础上，可以用一点在两帧之间的灰度变化来代表 I_t，用相对于其邻近点的灰度差异代表 I_x、I_y，那么，可以发现此约束方程有两个变量，那么如何求解呢？这就要用到之前的最后一个假设，即图像的每一小邻域中光流近似一致。因此，可以联立 n（n 为一个邻域内的总点数）个方程如下：

$$\begin{cases} I_{1x}u + I_{1y}v = -I_{1t} \\ I_{2x}u + I_{2y}v = -I_{2t} \\ \vdots \\ I_{nx}u + I_{ny}v = -I_{nt} \end{cases}$$

对于上面的方程组，可用最小二乘法求得最优解：

$$\begin{bmatrix} -I_{it} \end{bmatrix} = \begin{bmatrix} I_{ix} & I_{iy} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \left(\begin{bmatrix} I_{ix} & I_{iy} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} I_{ix} & I_{iy} \end{bmatrix} \right)^{-1} \begin{bmatrix} I_{ix} & I_{iy} \end{bmatrix} \begin{bmatrix} -I_{it} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} I_{ix}^2 & \sum_{i=1}^{n} I_{ix}I_{iy} \\ \sum_{i=1}^{n} I_{ix}I_{iy} & \sum_{i=1}^{n} I_{iy}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_{i=1}^{n} I_{ix}I_t \\ -\sum_{i=1}^{n} I_{iy}I_t \end{bmatrix}$$

以上就是 LK 光流算法的过程。此方法只能解决运动较小的情况。对于运动较大的情况，可采用金字塔分层的方式，缩小原图，使得 LK 算法继续适用，此处不再详细展开。

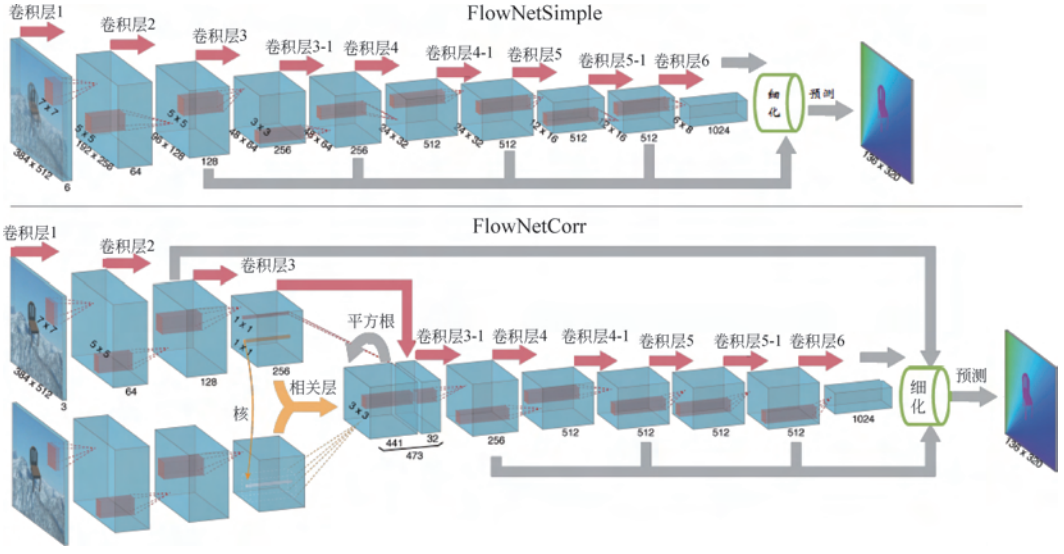
2. FlowNet

FlowNet 是一种基于深度学习的光流计算方法，如图 5-35 所示，给定一对时序相关的图片，通过卷积神经网络得到这两张图片之间的光流信息。由于光流问题与深度估计问题类似，其关键点都在于立体匹配

图 5-35　FlowNet 示意图。图片来源于 Dosovitskiy, et al., 2015

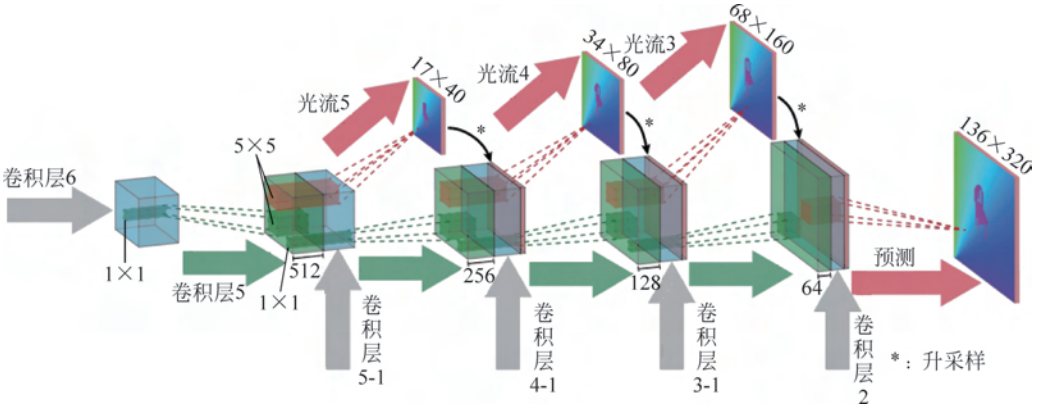
(stereo matching),只要能匹配一对图片中对应的点,那么不管是深度还是光流都可以比较容易地得到。所以,FlowNet 这种结构在双目深度估计中也有着较好的应用。

如图 5-36 所示,FlowNet 有两种实现方式:第一种是直接将成对的图片拼在一起,将输入通道数变成 6,通过一个卷积神经网络学习其特征和相关性,然后再经过细化网络,得到光流信息。第二种是将两张图分别经过一个参数共享的卷积神经网络提取特征,分别得到对应两张输入图的特征图,然后经过一个计算相关性的层,得到两张特征图上每一点在空间上的相关性,将此相关性矩阵与原特征图进行拼接,用于后面的预测。



■图 5-36 FlowNet 两种实现方式示意图。图片来源于 Dosovitskiy,et al.,2015

细化网络如图 5-37 所示,其作用是对结果进行上采样,得到高分辨率的结果图。这里用到了深度监督的方法,在上采样的每一步中,都会由一个卷积层输出该尺寸的结果并进行监督训练,并且,此结果进行上采样后会拼接到下一个维度的特征图中,参与更大尺寸的预测,这样会使得网络的输出越来越精细,最终得到一个分辨率较高的结果。



■图 5-37 细化网络示意图。图片来源于 Dosovitskiy,et al.,2015

5.6 基于 V2X 的道路环境感知技术

前序章节讲述的状态感知主要通过车载传感器对周边及本车环境信息进行采集和处理，包括交通状态感知和车身状态感知，与外界如道路的其他参与者不存在信息交互。而 V2X 网联通信是利用融合现代通信与网络技术，实现智能驾驶车辆与外界设施和设备之间的信息共享、互联互通和控制协同。本节将简单介绍 V2X 技术。

5.6.1 V2X 技术

1. 概述

V2X(Vehicle-to-Everything，车用无线通信技术)是将车辆与一切事物相连接的新一代信息通信技术。其中，V 代表车辆；X 代表任何与车交互信息的对象，主要包含车、交通路侧基础设施、人以及网络，分别采用以下缩写 V、I、P 和 N 表示。具体信息模式包括：车与车之间(Vehicle-to-Vehicle，V2V)、车与路侧基础设施(如红绿灯、交通摄像头和智能路牌等)之间(Vehicle-to-Infrastructure，V2I)、车与人之间(Vehicle-to-Pedestrian，V2P)、车与网络之间(Vehicle-to-Network，V2N)的交互。

V2X 将“人、车、路、云”等交通参与要素有机地联系在一起，不仅可以支撑车辆获得比单车感知更多的信息，促进自动驾驶技术创新和应用；还有利于构建一个智慧的交通体系，促进汽车和交通服务的新模式、新业态发展，对提高交通效率、节省资源、减少污染、降低事故发生率、改善交通管理具有重要意义。

1) V2X 通信优势

相比传统雷达，V2X 通信传感系统有以下几点优势。

(1) 覆盖面更广。

300～500m 的通信范围相比雷达探测范围要远得多，不仅是前方障碍物，而且身旁和身后的建筑物、车辆都会互相连接，大大拓展了驾驶员的视野范围，驾驶员能获得的信息也就更多，也更立体。因此，在前车刹车初期就能有效甄别，并进行提示，如果距离过近，系统会再次提示，对预判和规避危险也有足够的反应时间，避免出现跟车追尾的情况。

(2) 有效避免盲区。

由于所有物体都接入互联网，每个物体都会有单独的信号显示，因此即便是视野受阻，通过实时发送的信号也可以显示视野范围内看不到的物体状态，降低了盲区出现的概率，就充分避免了因盲区而导致的潜在伤害。

(3) 对于隐私信息的安全保护性更好。

由于这套系统将采用 5.9Hz 频段进行专项通信，相比传统通信技术更能确保安全性和私密性，如果通信协议及频道在各个国家都能够规范化，这套系统将变得像 SOS 救援频道一样成为社会公用资源。

2) V2X 通信的国内外发展进展

(1) 国外 V2X 通信的发展进展。

目前，这套 V2V 协议由通用、福特、克莱斯勒等厂商联合研发，除了美国汽车这三巨头

以外，丰田、日产、现代、起亚、大众、奔驰、马自达、斯巴鲁、菲亚特等车企也在协议名单内。2016 年 12 月 14 日，美国交通部发布了 V2V 的新法规，并进行 90 天的公示，法规强制要求新生产的轻型汽车安装 V2V 通信装置，这是一个里程碑式的进步。

美国交通部的新规中要求 V2V 装置的通信距离达到 300m，并且是 360°覆盖，远超摄像头的探测能力，其感知信息属于结构化信息，不存在误报的可能。根据美国高速公路安全管理局（NHTSA ）的研究，利用 V2X 技术，可以减少 80％的非伤亡事故，但这一切是以 100％的覆盖率为前提的。在此之前，如凯迪拉克等车企也曾经做过尝试但都因缺乏足够的覆盖率难以发挥作用，依靠强制性的法规驱动，V2X 普及的最大难题将得以有效解决。

高通发布新闻表示，将与奥迪、爱立信等公司进行蜂窝-V2X（Celluar-V2X）的测试合作，该测试符合由德国政府主导的项目组织——自动互联驾驶数字测试场的测试规范。在此之前，高通推出的基于其最新骁龙 X16 LTE Modem 的全新联网汽车参考平台，支持作为可选特性的专用短程通信（DSRC）和蜂窝-V2X。

（2）中国 V2X 通信的发展进展。

2016 年下半年，发改委连同交通部联合发布了《推进“互联网＋”便捷交通促进智能交通发展的实施方案》，明确提出“结合技术攻关和实验应用情况，推进制定人车路协同国家通信标准和设施设备接口规范，并开展专用无线频段分配工作”的标准制定工作。从目前的情况来看，LTE-V（长期演进技术-车辆通信）极有可能被确定为中国标准。5G 的推进对 V2X 是非常大的利好，因为 5G 标准本身就包含了 V2X，可以说 5G 的发展和无人驾驶的发展是相辅相成、互相促进的。

为了满足在商业应用上的高可靠性，越来越多的车企意识到在增强车辆能力的同时，需要将道路从对人友好改造为对车友好。从 2015 开始，中国所有的无人驾驶示范园区都在规划部署路侧系统（V2I ）。随着 5G 的时间表日渐清晰，更大范围的部署也让人非常期待。5G 的核心推动力来自物联网，而汽车可能是其中最大的单一应用，一辆无人车每天可以产生超过 1TB 的数据。目前，多个地图供应商正在积极准备用于无人驾驶的实时高精地图，以克服静态高精地图无法适应道路变化的难题，但之前受制于无线带宽，很难达到实用。5G 可提供高达 10Gb/s 的峰值速率，以及 1ms 的低延时性能，可以满足这样的需求。

2. V2V 技术

V2V 是指通过车载终端进行车辆间的通信。车载终端可以实时获取周围车辆的车速、位置、行车情况等信息，车辆间也可以构成一个互动的平台，实时交换文字、图片和视频等信息。将 V2V 技术应用于交通安全领域，能够提高交通的安全系数、减少交通事故、降低直接和非直接的经济损失，以及减少地面交通网络的拥塞。当前面车辆检测到障碍物或车祸等情况时，它将向周围发送碰撞警告信息，提醒后面的车辆潜在的危险。

3. V2I 技术

V2I 是指车载设备与路侧基础设施（如红绿灯和智能路牌等）进行通信。路侧基础设施也可以获取附近区域车辆的信息并发布各种实时信息。V2I 通信主要应用于道路危险状态提醒、限速提醒、信号灯提醒、滤波同行。

V2I 技术也是未来智能城市组成的一部分，但要普及该技术，更多的是考验城市的基础设施条件，同时对安全性也有严苛的要求，尤其是网络安全。如果 V2I 技术的系统被黑客

所利用，那么后果不堪设想。

4. V2P 技术

V2P 通过手机、智能穿戴设备（如智能手表）等实现车与行人信号交互，在根据车与人之间速度、位置等信号进行判断。有一定的碰撞隐患时，车辆通过仪表及蜂鸣器，手机通过图像及声音提示注意前方车辆或行人。V2P 通信主要应用于避免或减少交通事故等。行人检测系统可以在车辆、基础设施中或与行人本身一起实现，以向驾驶员、行人或两者提供警告。当车内警报系统变得越来越普遍（例如，盲点警告、前向碰撞警告）时，在车内警告路上有行人存在也是切实可行的。而对于路上的行人来说，最简单和最明显的行人警告系统则是手持设备，如手机、智能手表等。

现有的一些警告方式有：允许盲人或视力低下的行人的智能电话自动呼叫的应用程序；当信号交叉口的人行横道内的行人在公交车的预定路径中时，利用车内设施警告公交车驾驶员；当行人在红灯时横穿马路的警告，以及试图转弯的司机被警告在人行横道上有行人等。

5. V2N 技术

对于 V2X，V2N 允许在车辆和 V2X 管理系统以及 V2X 应用服务器之间进行广播和单播通信，通过使用蜂窝网络来实现。车辆能够收到有关道路上发生的交通事故的广播警报，或原计划路线上的拥挤或排队警告等。V2V 和 V2I 代表的都是近距离通信，而通过 V2N 技术可以实现远程数据传输。随着 5G 的到来，V2N 的能力会进一步加强，更有助于自动驾驶信息的获取与传输。

5.6.2　路侧感知技术

1. 概述

图 5-38 为典型路测感知技术方案示意图。

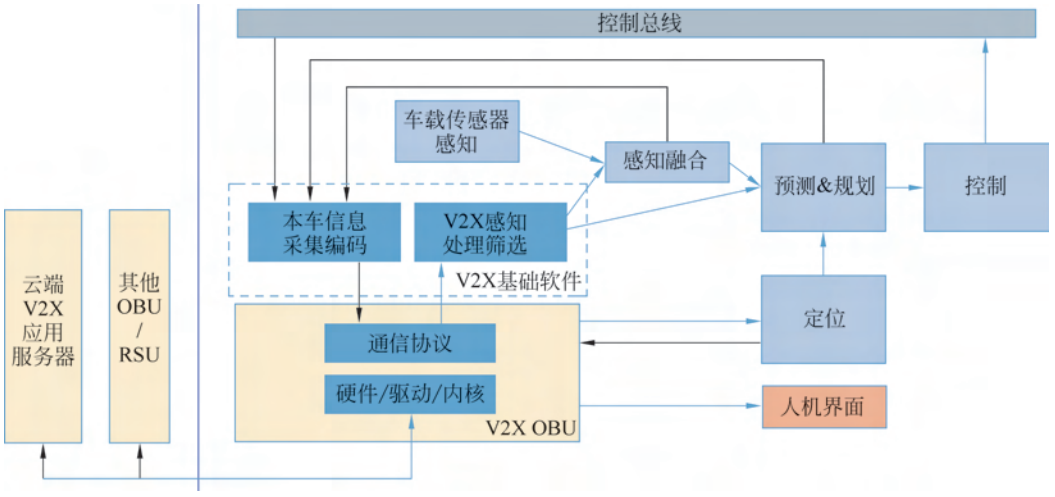


图 5-38　典型路侧感知技术方案示意图

百度宣布到 2018 年底百度将正式开源 Apollo 车路协同方案，向业界开放百度 Apollo 在车路协同领域的技术和服务，让自动驾驶进入“聪明的车”与“智能的路”相互协同的新阶

段，全面构筑“人-车-路”全域数据感知的智能交通系统，这是业内首个开源的车路协同方案。

Apollo 此次的车路协同开源，将在 Apollo 开放平台现有的四层开放技术框架的基础上，在软件、硬件、云端服务等层面增添或升级车路协同相关模块。在参考硬件层，Apollo 将增加车端以及路侧的参考硬件，用来完成自动驾驶车辆与路侧的信息传输与解析。在开源软件层，Apollo 对感知和决策规划模块进行了升级，能够完成 Apollo 系统车端对车路协同 V2X 相关信息的融合处理；同时提供可运行在车端及路侧参考硬件上的软件包，负责 V2X 信息的相关预处理工作。

在云服务层，开放智能路侧服务，提供自动驾驶所需的路侧感知预测等信息，同时开源路侧的感知预测等算法；并升级仿真服务能力，扩充在车路协同环境下的仿真场景。百度还宣布将基于全球领先的 Apollo 开放平台强大的生态能力，与大唐电信集团、千方科技、中国联通等产业链关键环节的代表性企业展开合作，全面整合汽车制造、交通基础设施设备制造和集成、通信、芯片、政府及高校等各界资源，共同发展车路协同系统。

百度还将与雄安新区合作，先试先行，进一步探索车路协同的交通发展新路径。同时，将与同济大学成立联合实验室，在无人车路网规划设计、交通流仿真等方面展开深度合作。

Apollo 拥有北京、雄安、硅谷等多样地区场景以及乘用车、无人小巴、无人物流车等多种车型；Apollo 在路侧感知传感器方案、路侧感知算法、车端感知融合算法、数据压缩与通信优化、V2X 终端硬件及软件、V2X 安全方面布局研发领先的车路协同全栈技术；Apollo 拥有的无人车队、开放道路无人车测试里程等一系列的场景数据积累，为百度布局车路协同、智能交通建设打下根基。

2. 车路协同技术

车路协同系统(Cooperative Vehicle Infrastructure System，CVIS)是基于无线通信、传感探测等技术获取车辆和道路信息，通过车-车、车-路通信实现信息交互和共享，从而实现车辆和路侧设施之间智能协同与协调，实现优化使用道路资源、提高交通安全、缓解拥堵的目标。近些年，智能汽车和无线通信技术的快速发展与应用，实现了车路协同技术在交通领域的发展。车路协同是智能交通系统(ITS)的重要子系统，也是欧、美、日等交通发达国家和地区的研究热点。

车路协同系统(CVIS)作为 ITS 的子系统，是将交通组成部分——人、车、路、环境——利用先进的科学技术(包括现代通信技术、检测感知技术以及互联网等)以实现信息交互的交通大环境；通过对全路段、全时间的交通动态信息采集与融合技术来提升车辆安全、道路通行能力以及智能化管理程度，达到加强道路交通安全、高效利用道路有限资源、提高道路通行效率与缓解道路拥堵的目标，形成安全、高效、环保、智能的交通环境。

系统是由路侧单元、车载单元、中心管理服务器、视频控制系统、信号控制系统几部分组成，各部分通过 DSRC、4G 网络、视频网、信号专网建立连接。其中，路侧单元(Road Side Unit，RSU)是可以检测自身状态信息、感知周围交通环境(包括交通流信息、道路几何特性、路面特殊事件交通信号控制器状态等信息)以及装配有无线通信模块和存储模块；车载单元(On Board Unit，OBU)主要是实现获取车辆状态信息、对车辆周围环境的感知(包括其他车辆、障碍物等)、安全预警和车载控制等功能，并能车-车、车-路通信，通过车载界面为驾驶员提供判断依据；中心管理服务器负责整个系统的通信、监控、下发数据与信息交互管

理等。

对交通控制方法的研究经历了几个阶段，主要是固定配时方法、感应控制方法、区域协调控制方法。区域协调控制方法更适用于现如今复杂多变的交通环境。现在，主流区域协调控制主要分为两大类：基于方案选择的交通信号控制和基于排队模型的交通信号控制。

基于方案选择的交通信号控制在投入运行之前，需要工程师制定相应的配时参数——交通量等级对应关系表，并将其事先存储于中央控制计算机内。将车辆检测器检测到的数据经过平滑处理得出交通量指数，中央控制计算机据此选择合适的配时参数组成配时方案；或者将配时方案与交通量指数的对应关系直接存储于中央控制计算机内部，控制层根据不同的交通量指数选择不同的方案。该种方法占用的 CPU 较少，处理更快。但其事先需要大量的交通调查，将配时参数与交通量指数对应起来，并且没有使用交通流模型，限制了方案的优化；检测器安装在停车线处，没有检测到后面车辆的到达情况，会影响相位差精度。

排队模型为基础的控制方法不需要事先存储任何交通配时参数、方案和事先制定各个参数与交通量指数的对应关系，而是通过实时检测到的交通流量和人工输入的优化器建模参数，经过推理、预测，整合为排队模型，得出相应结果，并通过优化器的优化机制，结合目标函数的优化，不断调整小步长，以适应交通流量的变化。它所采用的小步长，避免了方案较大跳动而引起路口短时间内的紊乱。但是建立交通模型需要采集大量的信息以及人工划分子区；阶段和放行顺序是固定不变的，不能自动改变；释放速率是采用固定值(饱和流率)，并没有按实际情况自动校准；断面检测只能获知车辆的存在、占有率、速度等低维数据，且检测数据对线圈设置位置有很高的要求，易受右转和路段进出口车辆的影响，因而车辆到达曲线可靠性略差。

5.7 红绿灯检测实验

红绿灯检测是自动驾驶中非常重要的任务之一，在无人车行驶过程中路口获取红绿灯的状态对无人车的感知和决策都是十分重要的信息，因此红绿灯检测应该能及时获得红绿灯的位置和颜色。目前，红绿灯检测主要通过摄像头数据进行输入，通过深度学习的方法获得需要的信息。

红绿灯检测本质上是一种目标检测，近年来卷积神经网络的快速发展使高精度、高准确度的目标检测成为了可能。自从使用 RCNN 方法以来，目标检测的速度和精度都不断提高，而对于红绿灯检测来说，卷积神经网络容易使红绿灯这样的小目标丢失语义信息，从而降低对小目标的检测效果，因此在本节中，将使用结合 FPN 的方法来提高对小目标的识别效果。

5.7.1 Apollo 红绿灯数据集

1. 数据集的作用

深度学习的方法是需要先用已标注的数据对模型进行训练。要想对红绿灯进行准确的

分类与检测，离不开大量的已标注数据。对于本实验，我们需要一个已标注好的红绿灯数据集，标注的信息包括图片中红绿灯的类别以及位置信息。

2. Apollo 数据集的介绍

Apollo 的数据平台开放了包括激光点云障碍物检测分类、红绿灯检测、Road Hackers、基于图像的障碍物检测分类、障碍物轨迹预测、场景解析等大量人工标注的数据集。我们可以在 http://data.apollo.auto/? locale=zh-cn&lang=en 上申请使用这些数据。

在本实验中，需要用到 Apollo 数据平台中的红绿灯检测数据集。红绿灯检测数据集包括了大量人工标注的行车场景图片，其中红绿灯的位置信息以及类别信息已经被标识出来，可以利用这些数据进行模型的训练和测试。

3. 数据集格式说明

打开数据集，可以看到 trainsets 和 testsets 两个文件夹，trainsets 文件夹里面的数据是训练集，testsets 文件夹里面的数据是测试集。

打开 trainsets 文件夹，可以看到 images 文件夹和 labels 文件夹，以及一个 list 文本文件。images 文件夹里面是训练集中所有的图片信息，图片如图 5-39 展示；labels 文件夹里面是所有图片的红绿灯信息。打开 list 文本文件，可以看到每一行对应着一张图片与其标签文件。

图 5-39　数据集图片展示

打开第一张图片与其标签文件，可以看到图片中有一个红灯，而其标签中有一行数字，第一个数字代表类别，其中 1 代表红灯，2 代表绿灯。后四个数字代表位置信息，分别代表了红绿灯矩形框的左上角、右下角的横纵坐标。

4. 数据的统计分析

在对数据集进行分析后可以发现，图片的分辨率为 1920×1080 像素，而大多数的红绿灯目标的大小长度在 60～210 像素，宽度在 20～70 像素，而比例则普遍为 1∶3。对数据集

的了解可以在设计方法时更加得心应手。

5.7.2 实验流程

1. 数据处理

首先将 Apollo 红绿灯数据集下载下来，然后处理成可以比较方便使用的形式。在本实验中，将其处理成 coco 数据集的格式。本数据集分划分为三个部分：训练、验证、测试。其中训练集有 82 张图片；验证集有 18 张图片；测试集有 100 张图片。

2. 调节参数

1）环境参数

首先在 tool/train.py 中可以调整使用 GPU 的编号：

```
os.environ["CUDA_VISIBLE_DEVICES"] = "0"
```

这行代码表示使用的 GPU 的编号为 0，读者也可以根据自己的实际情况进行调整。同时也可以修改其他一些参数。

如果仅使用一块 GPU，则：`cfg.gpus = 1`

如果想开启 GPU 中的设置加速计算，则：`torch.backends.cudnn.benchmark = True`

如果设置为不使用分布式，则：`distributed = False`

2）模型参数

可以在 configs/faster_rcnn_r50_1x.py 中调整模型的一些参数。在前文中已经对数据集的数据进行了一些统计分析，因此可以修改模型中的参数以使得训练效果更佳：

```
anchor_scales = [8]          # anchor 的大小
anchor_ratios = [0.33]       # anchor 的宽高比
```

同时，由于红绿灯任务的特点，目标之间几乎不存在同类别目标有重叠的情况，因此可以将非极大值抑制的阈值调到更低一些，使得过滤的标准更为严苛。

```
nms = dict(type = 'nms', iou_thr = 0.3)
```

读者也可以试试其他的参数，看看对结果有什么影响。

3. 训练及测试

在调整了参数之后，可以开始对模型的训练。首先在 config/faster_rcnn_r50_fpn_1x.py 中 91 行修改 data 的绝对路径（data 的绝对路径是指训练集的路径，在本项目中应为：本模型路径＋'data/coco/'）。然后输入以下命令进行训练：

```
./tools/dist_train.sh ./configs/faster_rcnn_r50_fpn_1x.py 1
```

后面两个参数分别是对应 faster_rcnn_r50_fpn_1x.py 的绝对路径和使用 GPU 的数量。

经测试在只使用一块 GPU 的情况下，大致需要 15min 的训练时间，如果使用多块 GPU 时间会更短。

本实验训练一共 12 个 epoch，每训练一个 epoch 将保存一个模型的数据，因此最后将保

存 12 个模型数据，我们应该使用最后一个进行测试。

测试：

```
python mytest.py
```

本文件夹中得到的 result.jpg 即为结果，图片中的目标上方标示了类别及置信度。结果如图 5-40 和图 5-41 所示。

图 5-40　检测结果展示 1

图 5-41　检测结果展示 2

5.8 本章小结

本章介绍了自动驾驶车辆环境感知与识别系统的组成，研究了车道线、红绿灯和障碍物检测方法，分析了各种检测方法的检测效果，对环境感知与识别系统当前面临的问题以及未来的发展趋势做了总结和展望。从障碍物的状态上，可以将障碍物的种类分为静止障碍和运动障碍，本章主要讨论静止障碍物和动态障碍物。用于障碍检测的传感器一般有视觉传感器、激光雷达、微波雷达、激光、声呐等。本章对常用的障碍检测方法如基于传统计算机视觉、基于深度学习以及基于激光雷达的障碍物检测技术进行了详细的介绍。

虽然目前智能驾驶车辆研究的主要任务是实现安全、智能、快速地行驶，但可以想象，在未来智能驾驶车辆还需要与更复杂的环境进行交互，因此，智能驾驶车辆的进步离不开环境感知与识别技术的支撑和发展。未来智能驾驶车辆的应用及智能交通系统的构建，将不断对环境感知与识别提出更高的需求。在憧憬未来的同时，我们更应该关注如何实现更为准确、可靠、全面的环境感知与识别技术。

参考文献

[1] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems. 2015: 91-99.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[3] REDMON J, FARHADI A. YOLO 9000: Better, Faster, Stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7263-7271.

[4] REDMON J, FARHADI A. YOLO v3: An Incremental Improvement[J]. arXiv: 1804.02767. 2018.

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]//European Conference on Computer Vision. Springer, Cham, 2016: 21-37.

[6] CAO G, XIE X, YANG W, et al. Feature-fused SSD: fast detection for small objects[C]//Ninth International Conference on Graphic and Image Processing (ICGIP 2017). International Society for Optics and Photonics, 2018, 10615: 106151E.

[7] LIN T Y, DOLLÁR P, GIRSHICK R B, et al. Feature Pyramid Networks for Object Detection [C]//CVPR. 2017, 1(2): 4.

[8] QI C R, SU H, MO K, et al. PointNet: Deep learning on point sets for 3D classification and segmentation[J]. Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, 1(2): 4.

[9] ZHOU Y, TUZEL O. Voxelnet: End-to-end learning for point cloud based 3D object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4490-4499.

[10] NEVEN D, BRABANDERE D B, GEORGOULIS S, et al. Towards End-to-End Lane Detection: an Instance Segmentation Approach[C]//IEEE Intelligent Vechicles Sysnposium(IV). IEEE. 2018: 286-291.

[11] PAN X, SHI J, LUO P, et al. Spatial As Deep: Spatial CNN for Traffic Scene Understanding[C]// Thirty-Second AAAI Confience on Artifical In telligence. 2018.

[12] 李亮，李锋林. 智能车辆导航中障碍物检测方法研究[J]. 电子科技，2017,30(09)162－164,168.

[13] 黄如林，梁华为，陈佳佳，等. 基于激光雷达的无人驾驶汽车动态障碍物检测、跟踪与识别方法[J]. 机器人，2016，38(4)：437-443.
[14] 吴毅华. 基于激光雷达回波信号的车道线检测方法研究[D]. 合肥：中国科学技术大学，2015.
[15] DOSOVITSKIY A，FISCHER P，ILG E，et al. Flownet：Learning optical flow with convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015：2758-2766.
[16] EIGEN D，PUHRSCH C，FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]//Advances in Neural Information Processing Systems. 2014：2366-2374.
[17] GODARD C，AODHA M O，BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：270-279.
[18] ZHOU T，BROWN M，SNAVELY N，et al. Unsupervised learning of depth and ego-motion from video[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：1851-1858.
[19] 黄海. 视频与雷达数据融合在围界入侵报警的应用探讨[J]. 智能建筑与智慧城市，2019(06)：37-39.
[20] 刘国荣. 基于图像的车道线检测与跟踪算法研究[D]. 长沙：湖南大学，2014.
[21] 高嵩，张博峰，陈超波，等. 一种基于双曲线模型的车道线检测算法[J]. 西安工业大学学报，2013，33(10)：840-844.
[22] 陈山枝，胡金玲，时岩，等. LTE-V2X 车联网技术、标准与应用[J]. 电信科学，2018，34(04)：1-11.
[23] 段续庭，田大新，王云鹏. 基于 V2X 通信网络的车辆协同定位增强方法[J]. 汽车工程，2018，40(08)：947-951.
[24] BING L，QI S，KAMEL A E. Improving the Intersection's Throughput using V2X Communication and Cooperative Adaptive Cruise Control[J]. Ifac Papersonline，2016，49(5)：359-364.
[25] RESS C，WIECKER M. V2X Communication for Road Safety and Efficiency[J]. Auto Tech Review，2016，5(2)：36-41.
[26] LE L，FESTAG A，BALDESSARI R，et al. V2X Communication and Intersection Safety 2 Related Work[M]//Advanced Microsystems for Automotive Applications 2009. Springer Berlin Heidelberg，2009.