# 目标检测

- ## 主流模型

  1. ### Faster R - CNN：

     作为两阶段目标检测模型的代表，其先利用区域建议网络（RPN）生成可能包含物体的候选区域。RPN 通过滑动窗口在特征图上生成一系列锚框，根据与真实物体框的匹配程度筛选出可能的候选区域。然后，这些候选区域进入后续网络进行分类和位置精修。此模型的优势在于能够生成高质量的候选区域，通过多步处理提高检测精度，适用于复杂场景下对各类目标的检测。

  2. ### YOLO 系列（如 YOLOv5、YOLOv7）：

     属于单阶段目标检测模型，直接在图像上进行回归预测。它将图像划分为多个网格，每个网格负责预测特定数量的边界框和类别概率。YOLO 系列模型速度极快，能够快速处理图像，实时输出检测结果，满足实时性要求高的辅助视障出行场景，例如实时告知视障人员周围的目标情况。

  ### 评估指标：

  1. mAP：模型在所有类别上的平均检测精度。
  2. FPS：模型的推理速度。
  3. Recall：模型召回率。
  4. Precision： 模型的精确率

- ## SOTA模型

  ```
  目前目标检测的 SOTA 模型如 Scaled - YOLOv4，在 YOLO 系列基础上进行了改进，通过对网络结构的优化和训练策略的调整，在保持较高推理速度的同时，进一步提升了检测精度。相较于主流的 YOLOv5 和 Faster R - CNN，Scaled - YOLOv4 在 mAP 指标上有显著提升，尤其在小目标检测上表现出色；在速度方面，虽然比 YOLOv5 略慢，但仍能满足大部分实时应用场景。在视障人员出行辅助中，Scaled - YOLOv4 能够更准确地检测出周围的小目标物体（如小型交通标志、宠物等），为视障人员提供更全面的环境信息。
  ```

- ## BDD100K（Berkeley DeepDrive 100K）

  ### 特点：

  1. 是目前最大规模、内容最具多样性的公开驾驶数据集之一，提供了10万张关键帧图片，涵盖多种天气和光照条件。
  2. 图像格式为 JPEG，标注信息为 JSON 文件，数据集中的GT框标签共有10个类别，包括Bus、Light、Sign、Person、Bike、Truck、Motor、Car、Train、Rider等。标注信息还包括源图像的URL、类别标签、大小（起始坐标、结束坐标、宽度和高度）、截断、遮挡和交通灯颜色等。

**选择理由：**

1. 包含行人、车辆、交通灯、交通标志等类别，适合室外环境
2. 数据丰富，多样性高，涵盖不同天气（晴天、多云、阴天、下雨、下雪、雾天）和光照条件，适合实际应用场景。

- **COCO**

**特点：**

1. COCO数据集包含数十万张图片，涵盖了80种常见物体类别。这些图片包含了丰富的上下文信息，能够很好地模拟真实世界中的应用场景。
2. 标注数据使用 JSON 文件存储，有 80 种预定义的目标类别，如 'person'（人）、'bicycle'（自行车）、'car'（汽车）等。

**选择理由：**

1. 丰富的标注信息能让模型学习到更全面的特征。
2. 提高模型在实际应用中对复杂场景的适应能力。

- **WiderPerson**

**特点：**

1. 用于行人检测的多样化数据集，它包含了从各种场景中精选的图像，不仅限于交通场景。该数据集选择了13,382张图像，并标注了大约40万个注释，这些注释包含了各种遮挡情况。数据集涵盖多种场景类型，包括城市街道、公园、广场等。
2. 图像为JPEG格式，标注数据为TXT文件，每个行人都有对应的边界框坐标。提供行人的遮挡情况，分为四类：完全可见-完全遮挡-部分可见-严重遮挡。

**选择理由：**

1. 提供行人边界框和遮挡标签，适合训练高精度的行人检测模型。
2. 是行人检测常用的基准数据集之一，具有多种算法和工具支持。

# 图像分割

- ## 主流模型

1. **U - Net：**

   经典的语义分割模型，采用编码器 - 解码器架构。编码器部分通过卷积层不断提取图像特征，降低特征图的分辨率；解码器部分则通过反卷积等操作恢复特征图的分辨率，并通过跳跃连接将编码器中对应的特征信息传递过来，从而有效利用上下文信息。这使得 U - Net 对小目标和细节的分割效果较好，在分割 Cityscapes 数据集中的道路、人行道等元素时，能准确划分出边界。

2. **DeepLab 系列（如 DeepLabv3+）：**

   采用空洞卷积来扩大感受野，同时结合了空间金字塔池化（ASPP）模块，能够更好地捕捉多尺度的上下文信息，在复杂城市场景的语义分割任务中表现出色，为视障人员提供更准确的场景语义理解，是复杂场景语义分割的最佳模型之一。

3. **SegNet：**

   于编码器 - 解码器结构，解码器部分通过保留编码器的最大池化索引来恢复特征图的空间分辨率，在盲道分割这类对空间信息要求较高的任务中表现良好，能够准确地分割出盲道区域。

### 评估指标：

1. IoU：对于盲道分割和城市场景元素分割任务，IoU 可直观衡量模型分割的准确性。
2. Pixel Accuracy： 反映模型对每个像素分类的准确性，可用于评估盲道分割等任务中模型对像素级别的分类能力。

- ## SOTA模型

```
如 Mask2Former，它将语义分割和实例分割任务统一在一个框架下，通过基于查询的方法进行分割预测。与主流模型相比，Mask2Former 在复杂场景下的分割精度有显著提升，尤其是在实例分割方面表现出色。在 Cityscapes 和 Mapillary Vistas 数据集上，Mask2Former 的平均 IoU 比 U - Net 和 DeepLabv3 + 更高，能够更准确地分割出不同类别的物体和实例。
```

- ## Cityscapes

### 特点：

1. Cityscapes 数据集主要收集了城市环境中的图像，焦点在于城市街道、建筑、人行道、车辆等场景。
2. 数据集中的图像分辨率为 2048x1024 像素，确保了细节的保留，适合多种视觉任务。图像格式为 PNG，标注信息为 JSON 文件，包括类别如：建筑（Building）、道路（Road）、交通标志（Traffic sign）、行人（Person）、车辆（Car）、天空（Sky）等。

### 选择理由：

1. 标签精细程度高，准确性好。
2. 记录了不同季节和天气条件下的街道场景。每张图像都配有详细的注释，包括语义标签和实例标签数据集还提供了相应的右立体视图、GPS坐标、车辆运动数据和外部温度等元数据

- ## Mapillary Vistas

### 特点：

1. 这个数据集是目前世界上最大、最多样化的街道级图像公开数据集，包含了25,000张高分辨率图像，这些图像被注释为66/124个对象类别，其中37/70类别是特定实例的标签。这些图像覆盖了全球六大洲，包括不同的天气、季节和一天中的不同时间，以及来自不同成像设备和摄影师的视角。
2. 图像格式为 JPEG，标注信息为 JSON 文件，包含 66 个语义类别和 28 个实例类别，如：交通灯、交通标志、人行道、盲道（无专门类别，但可通过人行道标注间接识别）消防栓、路灯、宠物。

### 选择理由：

1. 包含交通标志、交通灯、行人、车辆等多种类别，满足视障人员的室外环境需求。
2. 支持实例分割任务，可以区分同一类别的不同对象（如多辆汽车、多个行人）。
3. 数据集涵盖全球不同地区的街景，具有高度的多样性和真实性，适合实际应用场景。

# 深度估计

- ## 主流模型

  1. ### Monodepth 系列（如 Monodepth2）：

     基于单目图像进行深度估计的模型，通过无监督或自监督的方式学习深度信息，利用左右图像的视差以及图像重建损失等进行训练，在 KITTI 数据集上取得了较好的效果，能够为视障人员提供周围环境的大致深度感知，是单目图像深度估计的主流模型之一。

  2. ### DenseDepth：

     利用全卷积神经网络（FCN）来预测深度图，通过在大量图像上进行训练，学习到图像特征与深度之间的映射关系，帮助视障人员判断障碍物的远近。

  3. ### PSMNet：

     能够更好地适应不同大小物体和不同距离场景的深度估计。它对遮挡区域和纹理较少区域的深度估计有更好的效果，能够提供更准确的深度信息。

  ### 评估指标：

  1. MAE： MAE 越小，说明预测深度与真实深度越接近，能准确反映视障人员周围环境的深度信息。
  2. RMSE： RMSE 不仅考虑了误差的平均大小，还对较大误差给予更大权重，更能反映深度估计的整体误差情况。
  3. Log Scale Error： 对于深度估计中较大深度值的误差评估更为敏感，适用于评估远距离物体的深度估计准确性。

- ## SOTA模型

  ```
  当前深度估计的 SOTA 模型如 AdaBins，它提出了自适应分箱的方法来预测深度，在处理不同场景和不同深度范围的图像时具有更好的适应性。与主流模型相比，AdaBins 在 MAE、RMSE 和 Log Scale Error 等指标上都有明显改善，尤其在复杂场景和大深度范围的深度估计上表现出色。。
  ```

- ## KITTI

  ### 特点：

  1. 规模大，包含200个场景，约15000张图像。多传感器采集数据。
  2. 图像格式为PNG，点云数据BIN，标注数据TXT文件。适合多种任务-深度估计，通过激光雷达点云生成深度图，或以稀疏深度图形式提供。深度范围为 0.5m 到 80m。

  ### 选择理由：

  1. 激光雷达和相机数据的高精度对齐，适合深度估计任务。
  2. 数据集涵盖城市、乡村和高速公路等多种场景，具有更好的普适性。

- ## TP-Dataset

  ### 特点：

  1. TP-Dataset特别针对视障人士的需求，聚焦于环境中的深度信息，以帮助开发适用于视觉障碍者的导航和辅助技术。
  2. 原始图像数据，通常是 RGB 格式。存储标签信息的文件，通常为 CSV 或 JSON 格式。

  ### 选择理由：

  1. 数据集专为视障人士的需求设计，更符合研究的目标。
  2. 数据集包含了校园、街道、火车站、公交站、地铁、社区和医院等多种典型场景，总共有1391张包含盲道的图片。

- ## DIODE

  ### 特点：

  1. 包含 25,000 张高分辨率图像，涵盖室内和室外场景，包括家庭、办公室、街道、公园等。
  2. 数据格式为PNG，深度数据为PNG文件，室内：墙壁、地板、天花板、家具等。室外：天空、道路、植被、建筑物等。室内场景为 0.5m 到 10m，室外场景为 0.5m 到 80m。

  ### 选择理由：

  1. 使用激光扫描仪和结构光传感器生成高精度的深度图，适合训练深度估计模型。
  2. 数据集蕴含室内室外场景，适合多场景感知任务，多样性高。

- ## NYU Depth V2

  ### 特点：

  1. 包含 464 个室内场景的 RGB-D 数据，深度范围 0.5m 到 10m。
  2. PNG 图像 + MAT 文件标注（深度图）。

  ### 选择理由：

  1. 适合视障人员的室内导航需求。