

设计方案说明书

目录

第一章	概述	2
第二章	问题描述	3
2.1	任务和目标	3
2.2	需求概述	3
第三章	技术方案与实现	3
3.1	数据处理	3
3.2	模型设计	4
3.2.1	目标检测追踪模型设计	4
3.2.2	测距模型设计	5
3.2.2.1	单目视觉测距方法	6
3.2.2.2	单目视觉测距原理概述	7
3.2.2.3	单目视觉测距步骤	7
3.2.3	单目深度估计	10
3.2.3.1	单目深度估计方法概述	10
3.2.3.2	monodepth2 模型介绍	11
3.3	模型训练环境	12
3.3.1	硬件环境:	12
3.3.2	操作系统	12
3.4	模型检测效果	12
第四章	总结	13

第一章 概述

目标检测方面，使用 YOLO V5 进行目标检测。YOLO V5s 在对象检测方面非常出色，尤其是在高帧率下也具有较高的精度和较快的推理速度，为了使 YOLO V5s 模型能够提高道路中猫狗的物体检测追踪性能，调整了 Anchor 的尺寸及 LOU 值、使用进化超参数、缓存图像等以取得更快的训练速度。进行迁移学习后，模型检测速度更快、识别精度更高，FPS 达 332，mAP 达到 0.86。首先，进行对数据进行筛选、格式转换。其次，进行数据增强，其中包括左右翻转、随机裁剪 24%、亮度减少 80% 和 20%、对比度下调 50%、饱和度下调 40%、色调下调 50%。为了避免模型欠拟合，通过 kaggle 猫狗检测数据集 ([Dog and Cat Detection](#)) 对 yolov5 再次进行迁移，训练得到能够高效检测猫和狗的 YOLO V5s 模型，最终模型检测精度达到 94%。

在基于模型的坐标变换测距方面，我们选择了单目视觉测距和单目深度技术融合的方式，它相对于多目视觉测距技术具有成本低廉、系统安装简单等优点。借助 MATLAB Camera Calibration Toolbox 及 Opencv 对相机进行标定获取相机内外参数，通过 YOLO V5s 输出检测目标的矩形框位置，本文选择 Bounding box 中心及下边框中点作为目标测距关键点，预测出两个点的世界坐标后进行权重分配的到目标世界坐标。

在单目深度估计方面，我们使用 SfMLearner 自监督单目深度估计方法，该算法包括两个主要的网络分别是 Depth CNN 和 Pose CNN。其中 DepthCNN 网络用来预测深度，Pose CNN 网络用来预测位姿即 6 个自由度（3 个旋转和 3 个平移）。在 KITTI 数据集上进行训练，使用单目深度估计模型对猫狗图片进行深度估计，获取测试图像的深度信息，最后通过相机参数和几何关系去计算目标到相机源的位置信息。

以上两个测距算法通过多传感信息融合技术实现优势互补，再经过卡尔曼滤波修正提高了算法的抗干扰能力，取得最优距离，输出关注目标的类别和世界坐标，完成基于单目摄像的猫狗距离检测。算法总体构架如图 1。

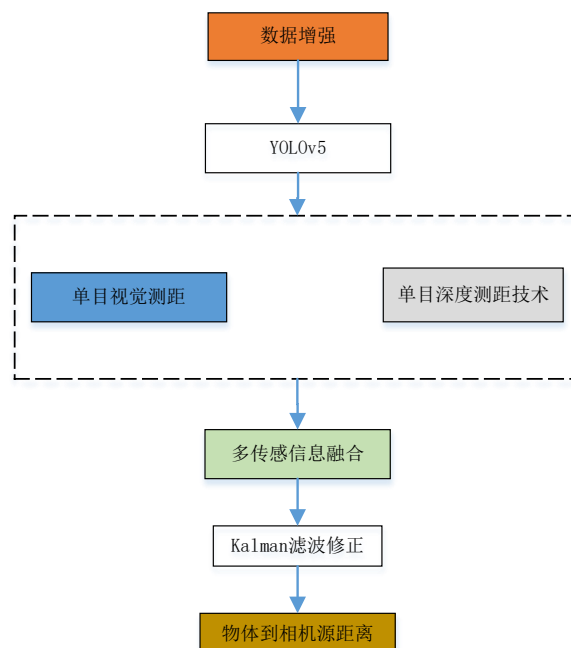


图 1 技术路线图

第二章 问题描述

车辆智能驾驶系统的环境感知中的一项重要任务，也是车辆后续路径规划、决策、控制的基础。智能驾驶中，道路中出现的猫狗会对汽车行驶造成严重危害，本方案针对道路交通环境下的猫狗为研究对象，进行目标跟踪并计算猫狗物体到汽车的距离，为车辆后续路径规划、决策、控制提供基础。

2.1 任务和目标

1. 利用提供的任务 5-领域一数据集，设计深度学习模型，训练得到能识别猫和狗的模型；
2. 完成猫狗目标跟踪、测距在内的完整检测功能。

2.2 需求概述

本方案主要针对路面中的猫狗物体为主要检测目标进行分析，首先设计针对猫狗物体的检测跟踪模型得到在较短时间情况下能够训练准确高效的猫狗物体目标追踪模型；其次，通过本方案测距算法进行测量猫狗物体到汽车的距离。

第三章 技术方案与实现

3.1 数据处理

本文基于 YOLO V5s 进行改进作为目标检测追踪模型，需要将原有数据与标签处理为 YOLO V5s 专用数据集与标签。首先，需要将“上汽杯”提供的 xml 数据集转换为 YOLO V5s 可用的.txt 格式。其次，为了使设计的目标检测模型对从不同环境获得的图像具有更高的鲁棒性对数据集进行数据增强处理如表 1，其中包括左右翻转、随机裁剪 24%、亮度减少 65%和 145%、对比度下调 50%、饱和度下调 40%、色调下调 50%。通过 kaggle 猫狗大战数据集（Cats vs. Dog）对 YOLO V5s 中的 BackBone 进行预训练，训练得到能够高效识别猫和狗的 YOLO V5s 模型，使网络对猫狗特征的具有较高的提取和识别能力。

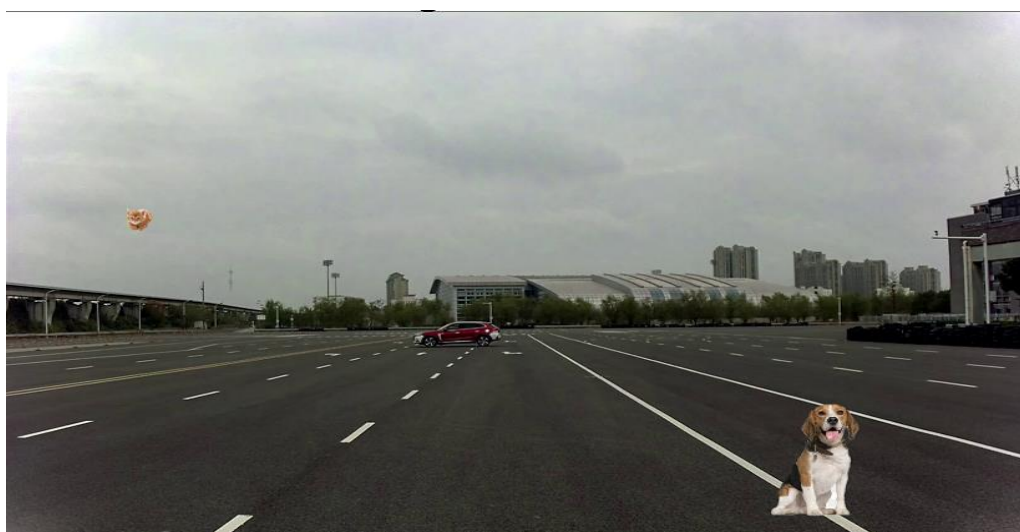

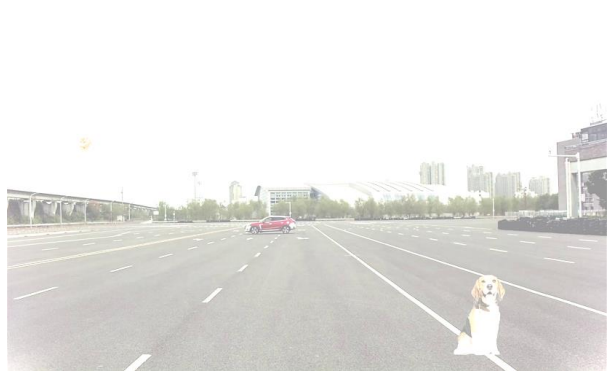
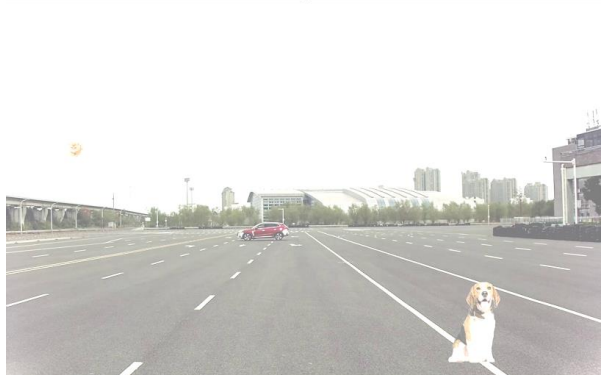


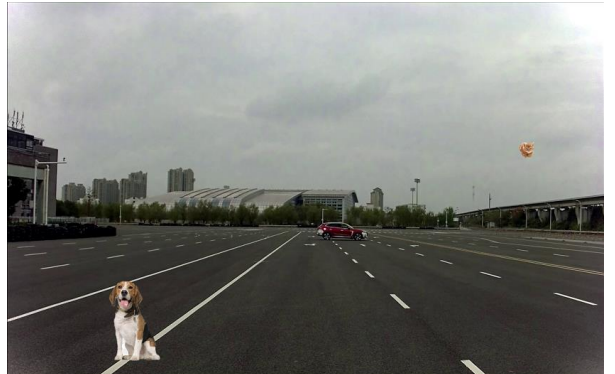


图 2 原始图像

表 1 数据增强图像

 <p>(a) 20%Brightness</p>	 <p>(b) 40%Brightness</p>
 <p>(c) 60%Brightness</p>	 <p>(d) 80%Brightness</p>
 <p>(e) Grayscale</p>	 <p>(f) flip horizontal</p>

3.2 模型设计

3.2.1 目标检测追踪模型设计

在车辆行驶过程中猫狗横穿马路会造成交通事故的发生，为了预防交通事故的发生，要求车辆能够快速检测到猫狗物体，并及时测量当前车辆与前方猫狗的距离范围并做出预警，对汽车安全行驶有非常重要的意义。如图 3 和表 2 所示目标检测模型 YOLO V5s 在众多模型之中脱颖而出，在保持高精度的情况下，极大地提升了检测速度。本文采用通用检测模型 YOLO V5s 对猫狗进行检测，首先修改 YOLOV5 的损失函数，通过 kaggle 猫狗大战数据集（Cats vs. Dog）对 YOLO V5s 中的 BackBone 进行预训练，得到能够高效识别猫和狗的 YOLO V5s 模型，使其成为猫狗专用检测器。

YOLO v5s 在 COCO 数据集上的训练结果如图 X 和表 X 所示。YOLO v5s 较其他模型相

比识别速度最快，同时也具有较高的平均精度（AP）和帧率精度（FPS）。

YOLO V5S 具有以下优点：

1. YOLO v5s 使用 Pytorch 框架，对用户非常友好，能够方便地训练自己的数据集。
2. 不仅易于配置环境，模型训练也非常快速，并且批处理推理产生实时结果。
3. 能够直接对单个图像，批处理图像，视频甚至网络摄像头端口输入进行有效推理。
4. YOLO V5S 具有高达 140FPS 的对象识别速度。YOLO V5S 的 SpeedGPU 为 2.1ms 速度快，同时 YOLO V5S 的模型大小只有 7.5 多兆，降低部署成本，有利于模型的快速部署。

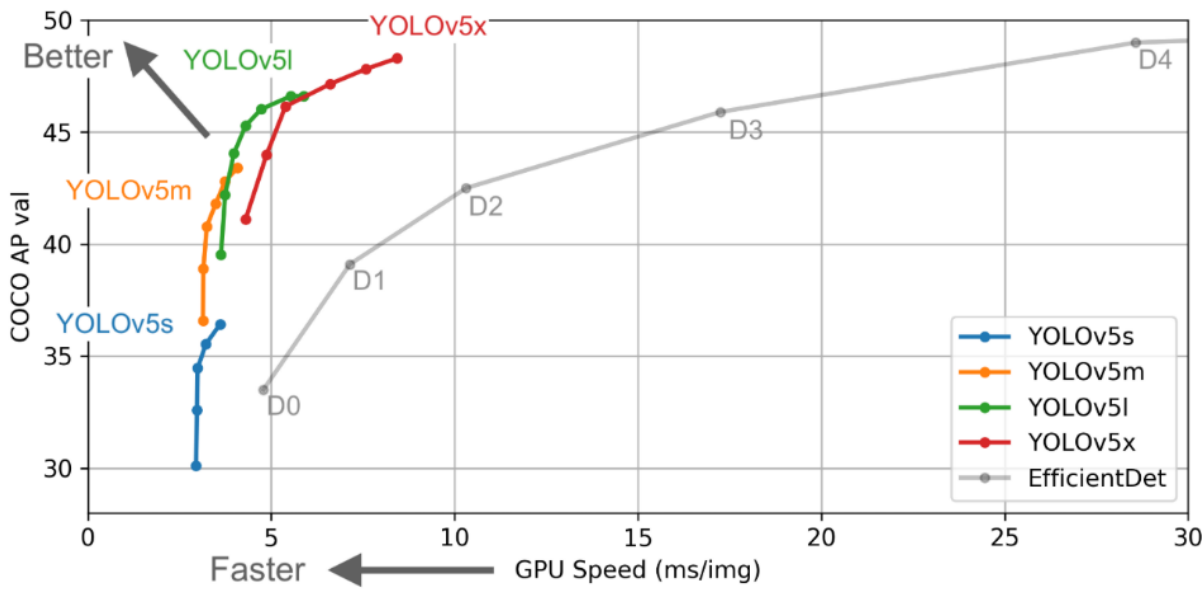


图 3 YOLO V5s 和其他目标检测模型在 COCO 数据集训练结果比较

表 2 YOLO V5 不同模型在 COCO 数据集训练结果对比

Model	APval	APtest	AP50	SpeedGPU	FPSGPU	FLOPS	params
YOLOv5s	36.6	36.6	55.8	2.1ms	476	13.2B	7.5M
YOLOv5m	43.4	43.4	62.4	3.0ms	333	39.4B	21.8M
YOLOv5l	46.6	46.7	65.4	3.9ms	256	88.4B	47.8M
YOLOv5x	48.4	48.4	66.9	6.1ms	164	166.4B	89.0M
YOLOv3-SPP	45.6	45.5	65.2	4.5ms	222	118.0B	63.0M

3.2.2 测距模型设计

视觉测距涉及范围广泛，如车辆、行人、树木、马路边缘、交通标志和路面标记等信息。视觉测距最早起源于摄影测量学，是把图像当作检测和传递信息的载体，从图像中获取被测对象的实际距离信息，在测距方面，我们选择了单目视觉测距技术与单目深度估计技术进行对比。对于单目视觉测距技术它相对于多目视觉测距技术具有成本低廉、系统安装简单、稳定性好等优点，而且不会出现复杂的图像匹配问题。深度估计是计算机视觉领域的一个基础性问题，其可以应用在机器人导航、增强现实、三维重建、自动驾驶等领域。而目前大部分深度估计都是基于二维 RGB 图像到 RGB-D 图像的转化估计，单目深度估计技术就是利用一张或者唯一视角下的 RGB 图像，估计图像中每个像素相对拍摄源的距离，通过获取相机参数利用相机几何关系可以求出来物体到拍摄相机的距离，从而进行准确测距。是由于双目图像需要利用立体匹配进行像素点对应和视差计算，所以计算复杂度也较高，尤其

是对于低纹理场景的匹配效果不好。而单目深度估计则相对成本更低，更容易普及。考虑到在汽车行驶过程中，需要实时地进行目标检测，针对传统目标检测与跟踪模型对比后选择 YOLO V5s 作为对象检测模型。YOLO V5s 在对象检测方面非常出色，尤其是在高帧率下也具有较高的精度和较快的推理速度。

3.2.2.1 单目视觉测距方法

检测流程如图 5 所示。首先利用数据集，训练得到能够识别猫和狗的 YOLO V5s 模型，然后将目标图像输入，识别出猫/狗类型，并且得到待测目标的矩形框位置，选择矩形框的下边中点作为目标关键点。接下来通过棋盘格标定法和 MATLAB Camera Calibration Toolbox 对相机进行标定，得到相机的内外参数，通过本方案目标测距算法测量目标物体与车之间的距离。最后进行误差修正，输出目标识别与测距的结果，完成检测。

本设计书针对给定的图像数据集，提供了一种基于单目视觉的目标测距方法。图 4 为单目视觉测距流程示意图。

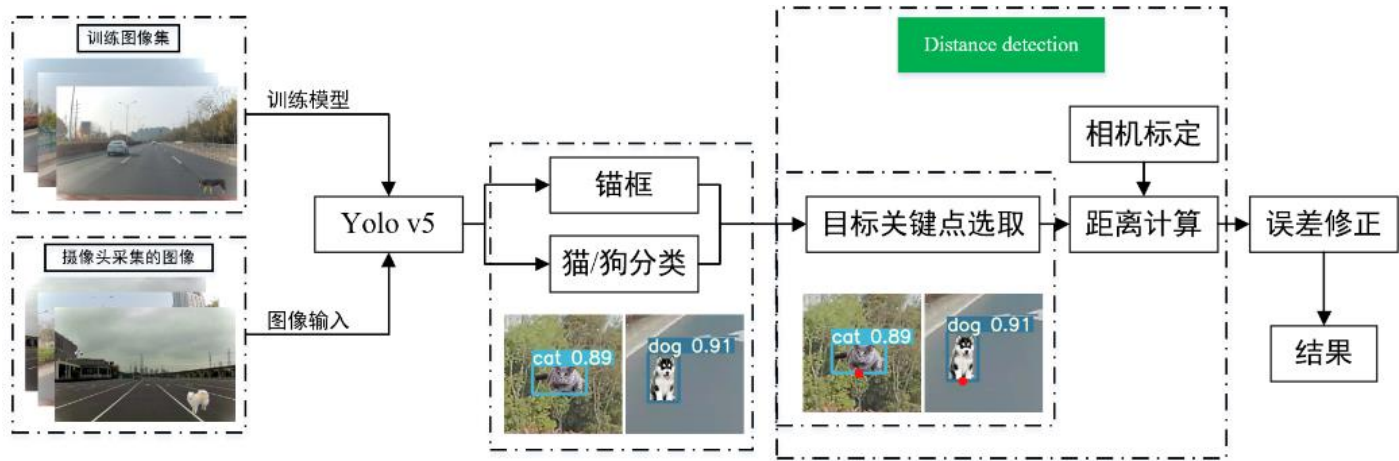


图 4 单目视觉测距流程示意图

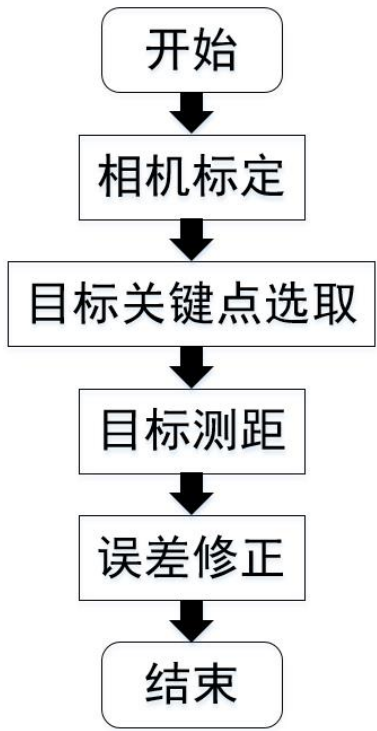


图 5 单目视觉测距流程示意图

3.2.2.2 单目视觉测距原理概述

单目视觉测距的原理，本质上就是世界坐标、相机坐标、图像坐标、像素坐标之间的转换。如图 6 所示， $O_w-X_wY_wZ_w$ 是世界坐标系， $O_c-X_cY_cZ_c$ 是相机坐标系，光心为原点， $o-xy$ 是图像坐标系，光心为图像中点， uv 是像素坐标系，原点为图像左上角。点 $P(X_w, Y_w, Z_w)$ 是世界坐标系中的一点，点 p 是点 P 在图像中的成像点，在图像坐标系中的坐标为 (x, y) ，在像素坐标系中的坐标为 (u, v) 。 f 是相机焦距，等于 o 与 O_c 的距离。

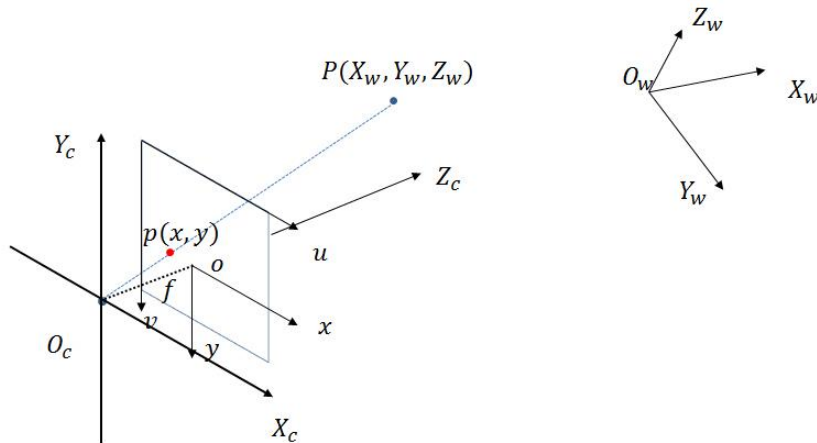


图 6 参考坐标系的关系

从世界坐标系到相机坐标系是刚体变换，对应旋转矩阵 R 和平移矢量 T ；从相机坐标系到图像坐标系是透视投影，对应透视投影矩阵 P ；从图像坐标系到像素坐标系是离散化，对应像素坐标和图像坐标的关系转换。最后得到了一个点从世界坐标系到像素坐标系的转换：

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_1 & a_2 & a_3 & x_0 \\ b_1 & b_2 & b_3 & y_0 \\ c_1 & c_2 & c_3 & z_0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}$$

3.2.2.3 单目视觉测距步骤

1. 相机标定

- 1) 打印一张用于相机标定的黑白相间的棋盘格，每个黑白方格的边长为 25mm，棋盘格尺寸为 10×7 ，把它贴在一个平面上，作为标定物；
- 2) 使用需要标定的相机捕获棋盘格图像，要求整张棋盘格在图像内，每个棋盘格边长不能少于 10 个像素，并且捕获从不同角度拍摄的 28 张图像；
- 3) 打开 MATLAB 软件中 Camera Calibration Toolbox 模块，点击 “add images”，然后输入模板的方格大小 25mm，选中步骤 1-2 中获取到的 28 张图像。如图 7 是相机标定示例图；

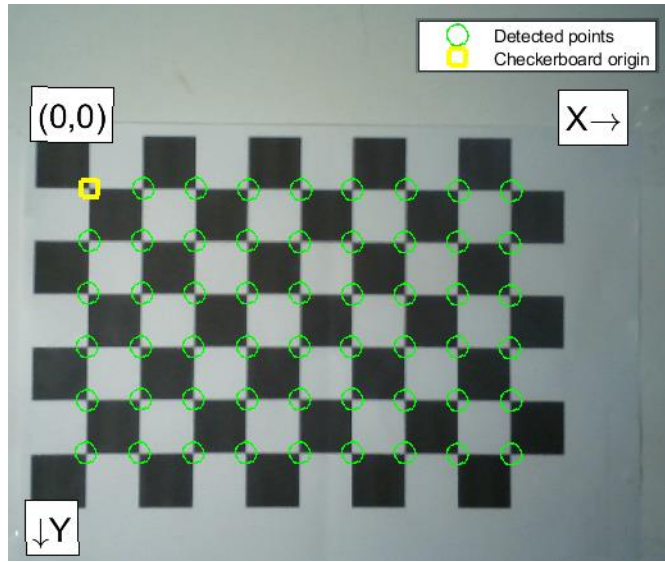


图 7 相机标定示例图

4) 添加完 28 张图像之后, 选择“calibrate”命令, 出现标定结果统计图, 点击保存 calibration. mat 文件, 相机标定完成, 用 MATLAB Camera Calibration Toolbox 进行相机标定的过程如图 8 所示;

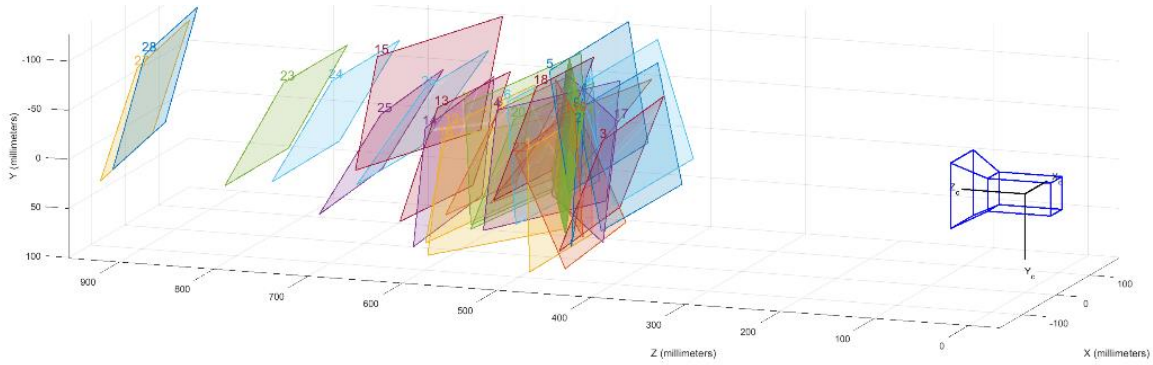


图 8 相机标定过程

5) 从 calibration. mat 中取出内参矩阵 $k = \begin{bmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$, 其中, f_x 、 f_y 分别是相

机水平方向和垂直方向相对单位像素的焦距, (u_0, v_0) 是光学中心坐标, 外参矩阵

$P = \begin{bmatrix} a_1 & a_2 & a_3 & x_0 \\ b_1 & b_2 & b_3 & y_0 \\ c_1 & c_2 & c_3 & z_0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, 其中, $R = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix}$ 、 $T = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix}$ 分别是世界坐标系到相机坐

标系的旋转和平移矩阵。

2. 目标测距点选取

1) 由已知的目标检测结果得到待测目标的矩形框位置 $(u \ v \ w \ h)$, 其中 $(u \ v)$ 表示矩形框在图像中左上角顶点的坐标值, $(w \ h)$ 表示矩形框的宽度和高度像素值;

2) 计算目标测距关键点 C 的像素坐标 $\left(u + \frac{1}{2}w, v + \frac{1}{2}h\right)$ 。目标关键点的选取如图 9 所

示。

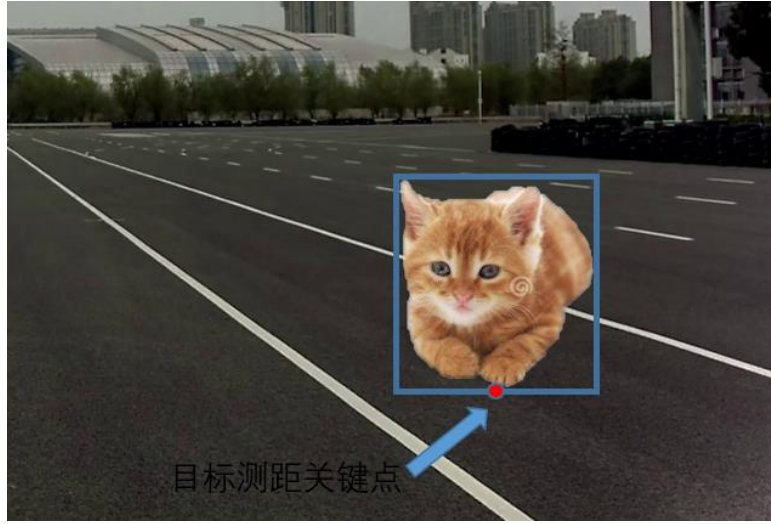


图9 目标关键点选取示例图

3) 校验测距关键点坐标合法性, 需要满足条件 $C = \{(x, y) | 0 < x < W, 0 < y < H\}$, 其中 W 是图像总的像素宽度, H 是图像总的像素高度。给定图像的 W 为 1920 像素, H 为 1208 像素。

3. 目标测距

1) 选取世界坐标系, 世界坐标系的坐标原点在相机正下方的水平路面上, x 轴方向为正前方, y 轴方向为正左方, z 轴方向为正上方, 符合右手定则;

2) 从 1 中得到相机参数的内参矩阵 K 和外参矩阵 P , 从 2 中得到待测目标的测距关键点 C 的像素坐标 (u_1, v_1) , 令测距关键点 C 的世界坐标为 (X_w, Y_w, Z_w) , 由于所计算的关键点 C 的世界坐标点位于水平地面上, 所以 $Z_w = 0$, 最后通过矩阵变换公式

$$s \cdot \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix} = K \cdot P \cdot \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \text{ 计算得到关键点 } C \text{ 的相关性未知尺度因子 } s;$$

3) 根据步骤 3-2 所求的相关性未知尺度因子 s , 关键点 C 的像素坐标以及相机参数的内参矩阵 K 和外参矩阵 P , 代入公式, 计算出像素坐标 C 点对应的世界坐标 $(X_1, Y_1, 0)$, 其中 X_1 即为待测目标关键点的纵向物理距离值, Y_1 即为待测目标关键点的横向物理距离值, 记为 $D_1(X_1, Y_1)$;

4) 由已知的目标检测结果获取到待测目标的矩形框位置 $(u \ v \ w \ h)$, 从中取得待测目标的像素宽度 w , 从 2 中得到测距关键点 C 的像素坐标, 计算出该点距离像素中心的横向像

素距离 $l = \left| u - \frac{1}{2} w \right|$, 由相似三角形原理得到公式 $\frac{f}{X} = \frac{w \cdot d_x}{W_1} = \frac{l \cdot d_x}{Y_2}$, 其中, 从 1 中

得到相机内参矩阵 K 中 f_x 的值, d_x 为相机横向像素单位大小, 相机焦距 $f = f_x d_x$, W_1 为目标实际物理宽度, X_2, Y_2 分别是需要求的目标纵向和横向物理距离, 记为 $D_2(X_2, Y_2)$ 。

4. 误差修正

- 1) 对计算结果 D_1 进行滤波处理，得到新的 D_1 ；
 - 2) 对 D_1 , D_2 进行卡尔曼滤波处理，得到滤波之后的结果 $(D_{cor1}, D_{cor2}) = \text{Kalman}(D_1, D_2)$ ；
 - 3) 分配权重并进行滤波处理，得到最终距离值 $D = \text{Kalman}(0.7D_{cor1} + 0.3D_{cor2})$ 。
- 相机均值误差如图 10 所示。

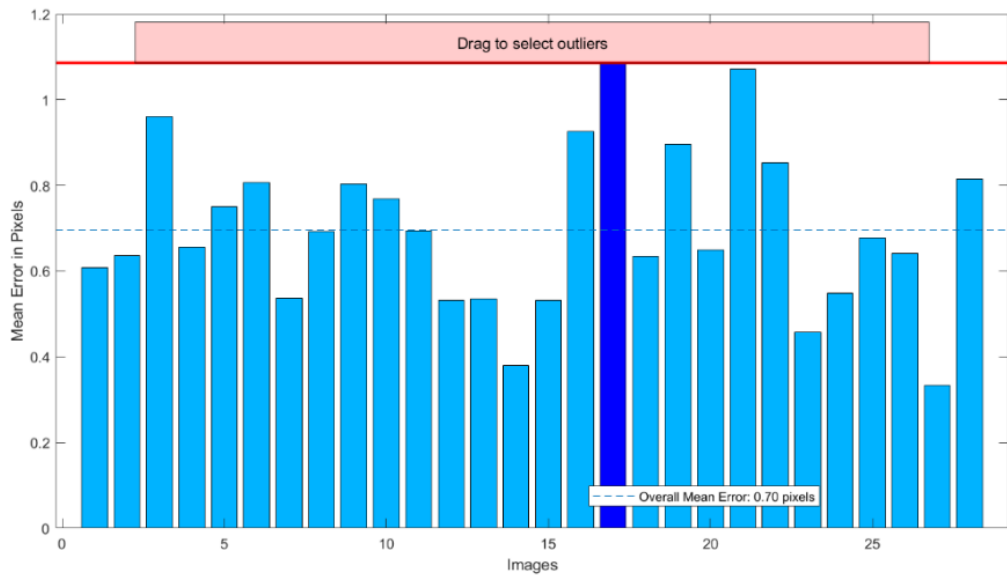


图 10 相机均值误差

3.2.3 单目深度估计

3.2.3.1 单目深度估计方法概述

单幅图像深度估计是计算机视觉中的经典问题，其可以应用在机器人导航、增强现实、三维重建、自动驾驶等领域。按照数学模型的不同，单目深度估计方法可分为基于传统机器学习的方法与基于深度学习的方法。基于深度学习的方法，又可分为有监督方法和无监督方法，有监督学习方法要求每幅 RGB 图像都有其对应的深度信息标签，因此受限于训练集场景。并且深度信息标签需要用深度相机或激光雷达来采集，但是深度相机范围受限，激光雷达成本昂贵，所以无监督估计方法的成本更低。因此不用深度标签的无监督估计方法是研究趋势。Monodepth2 是一种效果非常不错的自监督单目深度估计方法，它使用内在几何关系进行监督学习，同时它通过 pytorch 实现，对用户非常友好。所以我们选择 Monodepth2 作为单目深度估计模型。如图 11 所示，是 monodepth2 与其它几种单目深度估计方法在 KITTI 数据集上的对比，可见 monodepth2 产生了最清晰的深度图。

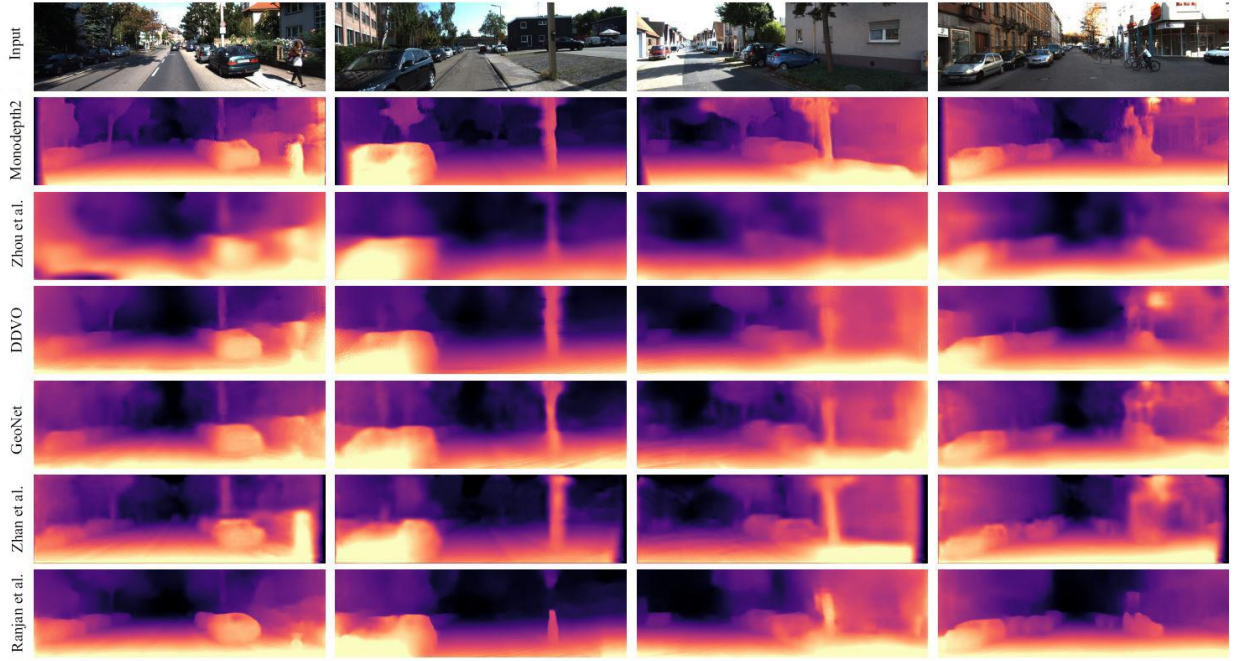


图 11 Monodepth2 与其它方法在 KITTI 数据集上产生的深度图的对比

3.2.3.2 monodepth2 模型介绍

monodepth2 方法使用深度估计和姿态估计网络的组合来预测单帧图像中的深度。它通过在一系列运动的图像序列上，训练一个建立在自监督损失函数上的架构来实现，这一架构包括两个网络，一个用来在单目图像上预测深度，另一个在运动图像之间预测姿态。此方法不需要标注训练数据集，相反，它使用图像序列中的连续时间帧和姿态的重投影关系来进行训练。

Monodepth2 模型的训练过程如图 12 所示。首先通过标准的、完全卷积的 U-Net 深度网络，根据颜色预测目标图像中的每个像素的深度值；然后通过位姿网络，将前后两帧彩色图像作为输入，来预测目标图像的单个 6 自由度相对位姿或旋转与平移参数；最后将得到的位姿矩阵投影到具有固定内参矩阵 K 的摄像机中，以获取重建的目标图像，计算每个像素值的重投影误差损失函数，最小化这个损失函数，以减少目标图像和重建的目标图像之间的差异。重投影误差的计算如下所示：

$$L_p = \sum_{t'} pe(I_t, I_{t' \rightarrow t})$$

$$I_{t' \rightarrow t} = I_{t'} \langle \text{proj}(D_t, T_{t \rightarrow t'}, K) \rangle$$

$$pe(I_a, I_b) = \frac{\alpha}{2} (1 - \text{SSIM}(I_a, I_b)) + (1 - \alpha) \|I_a - I_b\|_1$$

其中， pe 是重投影误差， D_t 是预测的深度图， $\text{proj}()$ 是投影图像的 2D 坐标， $T_{t \rightarrow t'}$ 是变换映射， K 是相机的内部参数， $\langle \rangle$ 是双线性采样算子， I_t 是目标图像， $I_{t'}$ 是相邻图像。

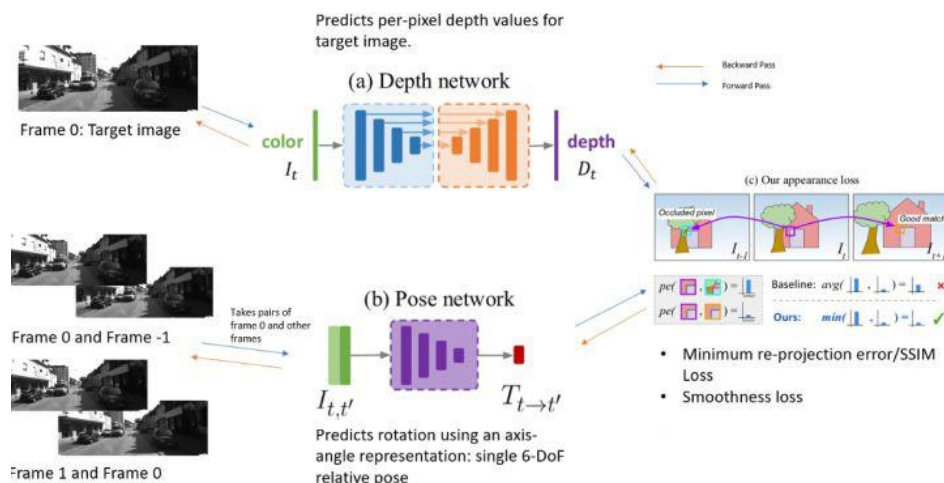


图 12 Monodepth2 模型训练过程

Monodepth2 单目深度估计模型具有以下特点：

1. 计算并最小化每个像素的重投影损失，以减少目标图像和重建的目标图像之间的差异。
2. 引入自动遮蔽损失，以忽略混乱的、固定的像素。
3. 全分辨率多尺度采样方法。该模型可以用单目视频数据、立体数据或混合的单目和立体数据进行训练。

3.3 模型训练环境

3.3.1 硬件环境：

CPU：Intel(R) Core(TM) i7-9700F CPU @ 3.00GHz

GPU：NVIDIA GeForce GTX 1660

3.3.2 操作系统

版本 Windows 10 家庭中文版

版本号 20H2

安装日期 2020/6/26

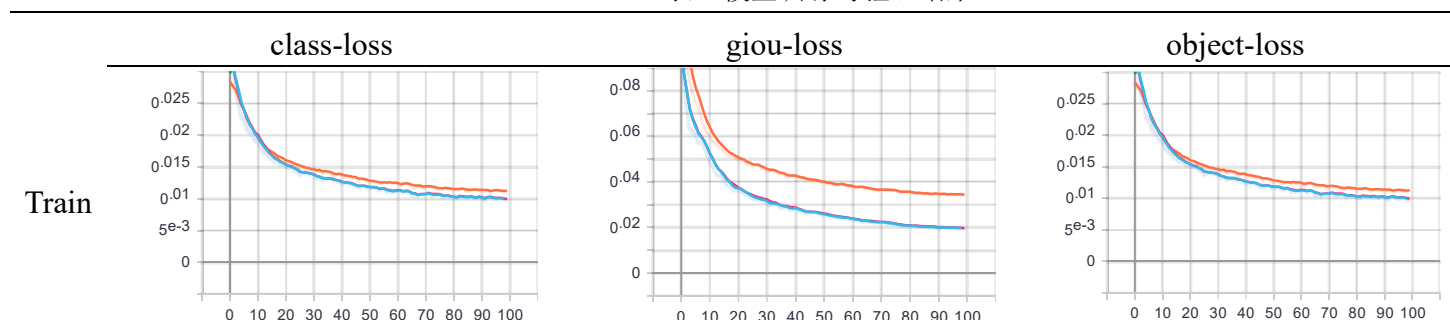
操作系统版本 19042.421

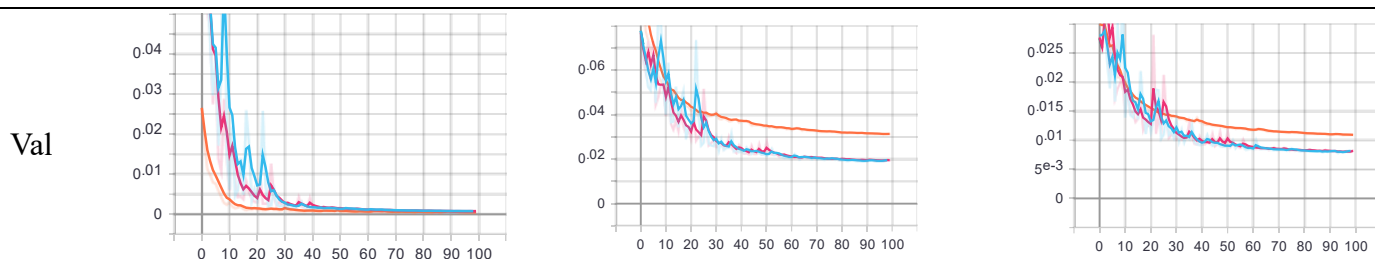
体验 Windows Feature Experience Pack 120.2212.31.0

3.4 模型检测效果

模型训练与验证结果如表 3 所示，class 损失值，giou 损失值，object 损失值。

表 3 模型训练与验证结果





模型训练效果如图 13 所示，mAP 为多个类别物体检测平均精确率的平均值，Precision 为识别的精确率，Recall 为识别的召回率。从图中可以看出，随着训练轮数的增加，识别猫、狗和车的 mAP 值持续增加，当训练轮数大于 25 的时候，mAP 值达到 0.86 左右，趋于稳定；随着训练轮数的增加，识别的精确率持续增加，当训练轮数大于 40 的时候，识别的精确率达到 0.7 左右，趋于稳定；当训练轮数小于 30 的时候，召回率震动比较剧烈，当训练轮数大于 30 的时候，召回率区域缓和，并稳定在 0.9 左右。

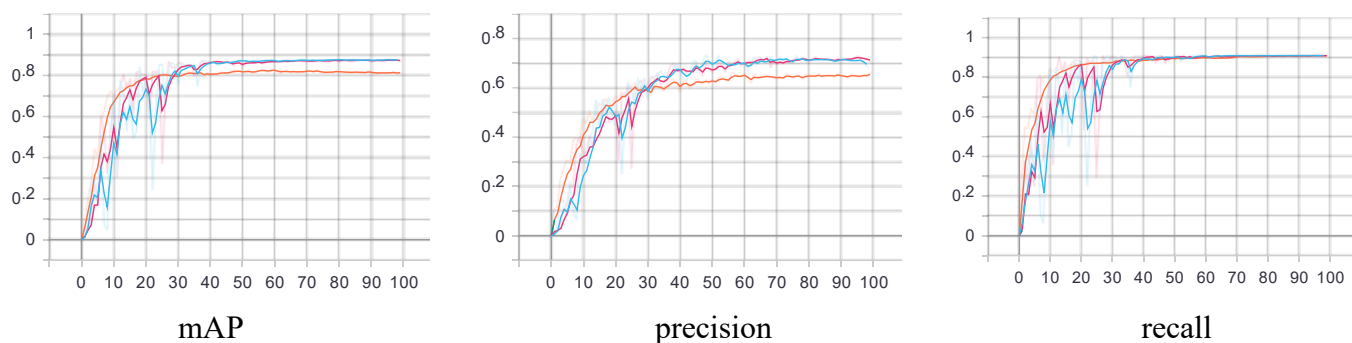


图 13 模型训练效果

第四章 总结

通过猫狗测距问题的研究发现，通过传统的相机模型的坐标变换方法鲁棒性和精度差，这些缺点限制了它在实车的应用；而基于深度学习的单目深度估计是目前可靠性最高的方法，前视摄像头采集的图像通过单目深度估计能够重建三维结构信息，对机器人自主导航、抓取等任务有重要意义，其主要面临的问题是数据与计算力的需求。通过两种方法的结合是本解决方案的一个创新性构想。

由于时间及计算力有限，本方案还需进一步优化。需要更多的猫狗数据进行训练，解决模型欠拟合问题；建立更精确的坐标转换模型来减小基于相机模型的测距误差；需要更多的深度估计数据及予以分割算法来提高单目深度估计的精度。