# 基于 SurroundDepth 和 nuscenes 数据集的监督学习改进方法

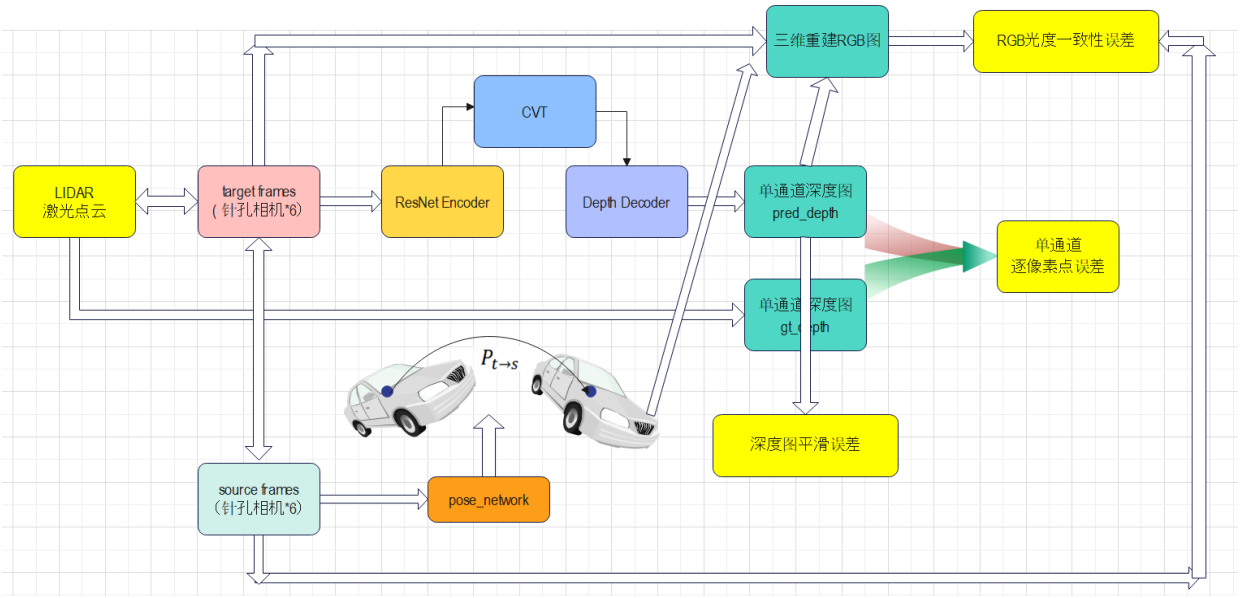
沈瑞淇

SurroundDepth 的原理部分已于第一部分讲述，这里不再重复。
本单元的核心为修改 SurroundDepth 论文中的预训练方式，将 SfM 自监督的部分改为监督学习，第二阶段仍为自监督学习不变。

**一．LIDAR 生成深度图的监督学习原理，与完整流程图**

监督阶段的核心原理为：由激光雷达 LIDAR 生成单通道深度图，直接将它作为 gt_depth 提供给网络进行预训练，使网络具备感知物体在世界中尺度的基本能力，达到和 SfM 类似甚至更好的效果。第二阶段仍为自监督学习，使用 pose_net 来预测车辆的自运动 ego_motion，自监督使用的损失函数仍为 RGB 三维重建损失与深度图平滑损失。

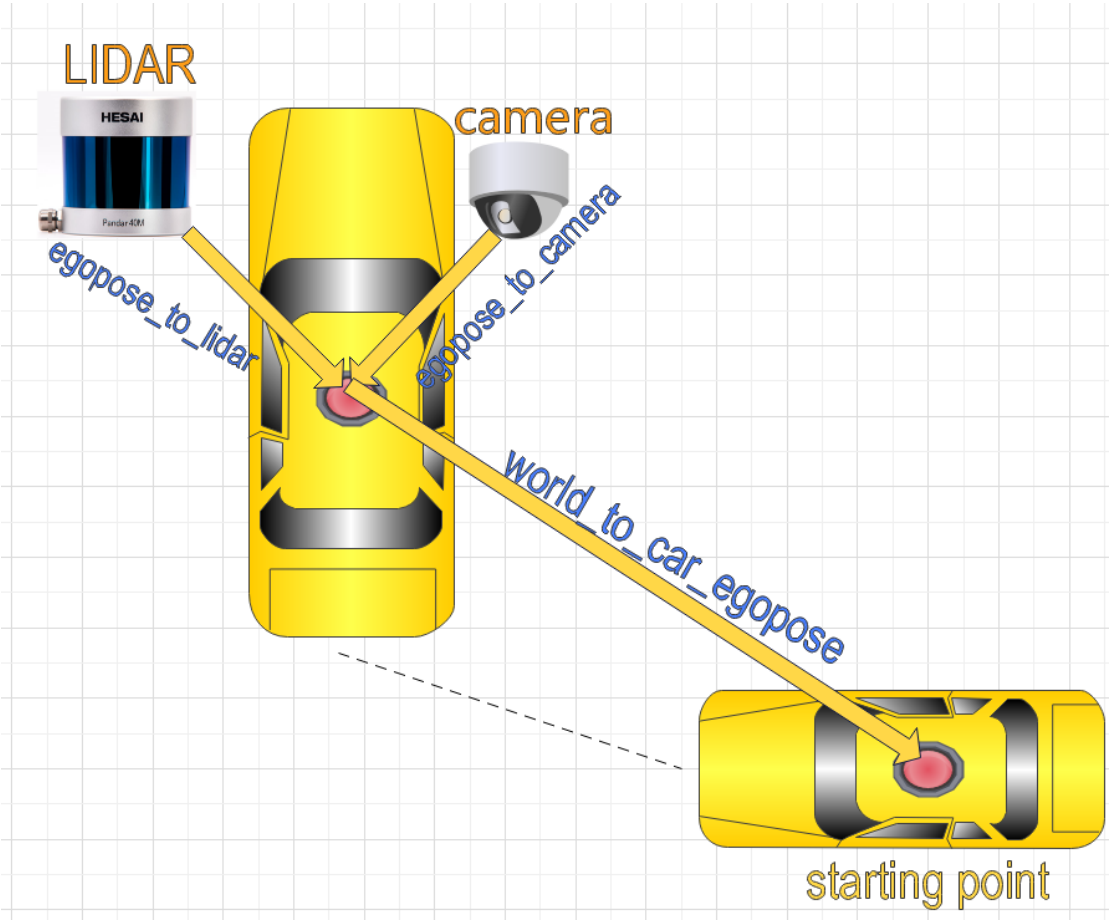
完整的两步训练流程如下所示：




LIDAR 生成稀疏深度图的原理：
LIDAR 直接获得的是三维空间的激光点云，当发出的激光与物体相交时，一部分激光会被反射回来，激光雷达接收来自各个方向的反射信号，从而能大致描述物体在空间的形状，以及距离激光束的距离。
在常见的自动驾驶任务中，由于传感器数量多，一般选择车中心点来表征自动驾驶汽车的实时位置。因此，LIDAR 激光点云需要通过以下步骤投影到任意的相机坐标系：


注意，尽管是同状态的 LIDAR 激光点云与相机坐标系，时间戳也不完全一样，会相差毫秒级别，这也是为什么不能够直接将三维点云投影到相机坐标系，而是必须由世界坐标系作为中

转站。激光点云的投影示意图如下：




图中，egopose_to_lidar 是 t1 时刻，egopose_to_camera 是 t2 时刻，而 world_to_car_egopose 严格来说，应该是两条线，即 t1 和 t2 两个时刻的车辆自运动坐标系与世界坐标系之间的转换关系。

有了以上的 lidar_to_camera 的转换关系后，对于激光雷达三维点云，可以将其投影到同状态的任意相机坐标系中，随后再通过该相机的内参（包括等效焦距和偏移点），投影到二维图像坐标系。

除此之外，任何投影点必须满足如下两个条件：1）深度大于 0，2）必须在图像长宽范围内我们用特定的掩码来过滤掉不满足要求的深度投影点，得到最终的稀疏深度图。

**二．实验代码文件用途解释**

代码文件夹路径：**/datab1/ruiqi_shen/bp/SurroundDepth**，conda 环境为 surrounddepth2，重点代码文件解释如下：

**export_gt_depth_nusc.py** 用来从 LIDAR 三维点云，生成稀疏深度图（为所有图像，包括训练和测试集）

**sift_nusc.py** 用来提取 CASIA 针孔图像的 sift 特征点，边缘阈值和对比度阈值维持不变，分别为 8 和 0.01

**match_nusc.py** .匹配 sift 特征点，每个视角的图像和左右两个相邻视角的图像做匹配

**nusc_dataset.py** 建立在 mono_dataset 基础上，建立完整的数据集，包含 RGB 图像信息，LIDAR 生成的深度信息，特征点匹配信息，内外参信息，运动姿态信息等。

**runer.py** 核心训练和验证函数，规定了训练超参数，数据集和网络模型加载，训练（损失函

数类型和定义），验证方法等
**run.py** 主函数，执行即可
**nusc_supervised_pretrain.txt** 第一阶段预训练（监督学习）配置参数
**nusc_supervised.txt** 第二阶段正式训练配置参数
**所有的 nuscenes 实验数据**，都在/datab1/ruiqi_shen/bp/SurroundDepth/data/nuscenes 下，包括原始数据集（6 个视角的图像和激光雷达点云），sift 特征点，特征点匹配结果，由 LIDAR 点云生成的稀疏深度图等

**三．实验结果**
这里将预训练阶段为 sfm 和 LIDAR 深度图监督学习的两种训练方式的点数与可视化效果放在一起展示：（当然都完成了第二阶段的正式训练）
指标参数如下：

$$\text{Abs Rel: } \frac{1}{|T|}\sum_{d\in T}|d-d^*|/d^*$$

$$\text{Sq Rel: } \frac{1}{|T|}\sum_{d\in T}||d-d^*||^2/d^*$$

绝对相对误差 Abs Rel，和平方相对误差 Square Rel，都是越接近 0 越好。
Median 之比：真实深度图的深度值中位数，与预测深度图的深度值中位数之比，越接近 1 越好（说明两幅图相似），表达式如下：

$$\frac{median(D^*)}{median(D)}$$

| | median 之比 | Abs Rel | Sq Rel |
|---|---|---|---|
| 预训练为 SfM | 1.0745 | 0.452 | 13.822 |
| 预训练为监督学习 | 1.5859 | 0.662 | 17.044 |

预训练为 SfM 时，按照要求训练了 5+2（即预训练 5 个 epoch，正式训练 2 个 epoch），点数最好时为正式训练第 1 个 epoch 结束时，指标如上，可视化视频也利用的这个模型。
预训练为 LIDAR 深度图监督学习时，训练了 5+4（即预训练 5 个 epoch，正式训练 4 个 epoch），点数最好时为正式训练第 3 个 epoch 结束时，指标如上。可视化时，发现正式训练第 1 个 epoch 结束时效果最好，因此可视化视频由此模型制作。
理论上，监督学习应该取得比 SfM 更好的点数，目前认为现今结果来源于未对 LIDAR 生成的原始激光点云做去噪声的过滤处理，否则来源于设备噪声和环境干扰的离群点会影响深度图的质量。可以考虑用滤波算法去除它们。（即 LIDAR 生成的稀疏深度图存在波纹效应，我们也将问题发邮件询问原作者，他也作出如上判断）。

可视化视频已经放在了附录中，它是由 nuscenes 数据集的 6019 组验证集数据（每组对应 6 个视角的 6 张图像），逐张通过训练好的模型做 validation 生成深度图，转为热力图表示，并逐帧拼接而成的视频（每个视角对应一个视频）。
其中，对应监督学习的视频命名为 nusc_supervised_0/1/2/3/4/5.mp4，0-5 分别对应自动驾驶汽车的 6 个针孔相机。对应 SfM 的视频命名为 nusc_sfm_0.mp4，因为我选择了第 0 个视角（即前视视角）来生成视频，用来做效果对比。
通过对比前视角（CAMERA_FRONT）的可视化视频可以发现，监督学习生成的模型虽然点数较低，但是整体可视化效果更好，尤其是车辆本身更为突出，详情请观看视频。