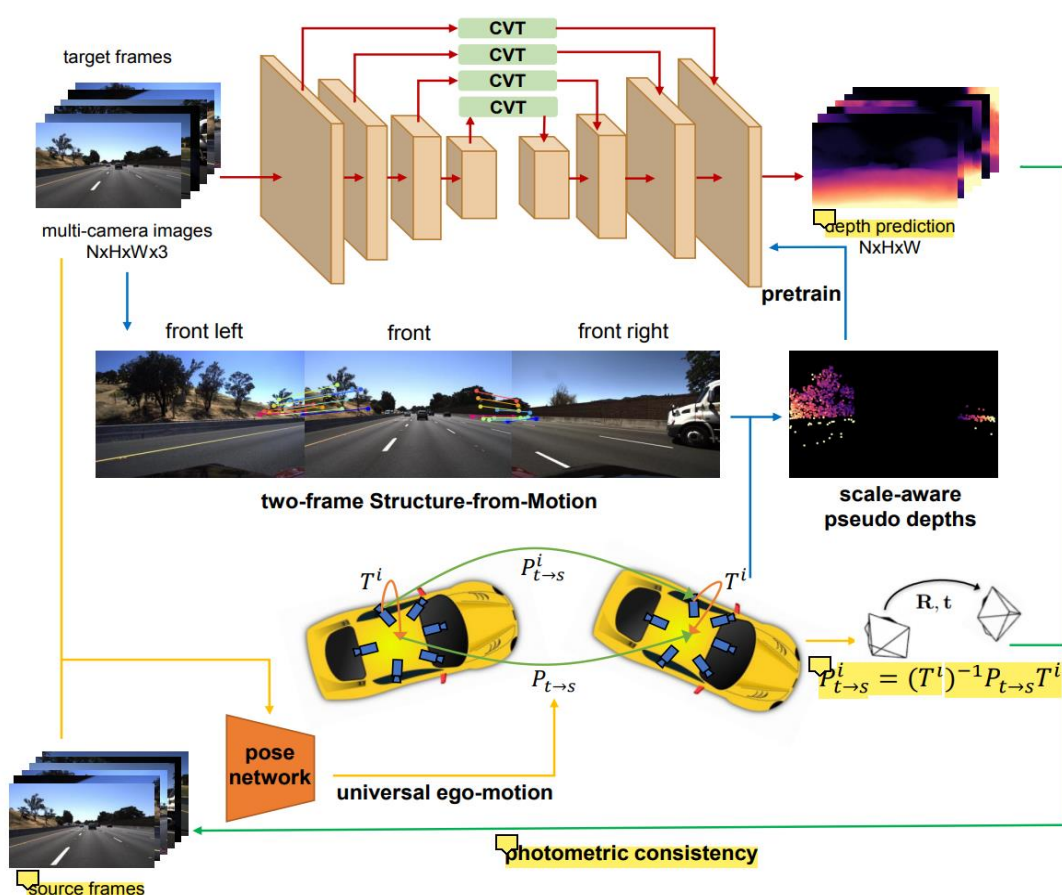


SurroundDepth 论文内容摘要

沈瑞淇

论文内容摘要

SurroundDepth: Entangling Surrounding Views for Self-Supervised Multi-Camera Depth Estimation



论文提出了一种自监督多视角深度估计的方法，通过融合相邻视角的图像信息，使得模型更好地进行深度估计。

具体来说，预训练使用 SfM(运动结构恢复)进行三维重建，通过匹配同一时刻相邻视角图像之间的 Sift 特征，并作为预训练真值更新参数，辅助模型大致判断出物体在真实世界中的尺度。正式训练时（输入为 target frames），一路网络预测深度，另一路通过多视角融合预测整车的自运动，并投影至每个相机，预测外参。深度和外参将重建三维图像，并与 source frames 时刻的图像计算 loss

论文使用 6 个针孔相机做深度估计（即每次输入为同时刻 6 个视角的图像），与本组的自动驾驶汽车相同，此外预测得到的深度图与输入图像是一对一的关系，而不是六对一。

我将从以下三个部分来详细介绍论文：1) SfM 预训练，2) 正式训练，3) 损失函数与掩码

一 . SfM 预训练

SfM 预训练的目的是通过相机外参来感知物体在真实世界中的尺度。每次输入同一时刻来自 6 个视角的 6 幅图像，提取出 6 张图像的 sift 特征点。由于每幅图像和其左右相邻视角的图像有视野上的重叠，因此对每幅图像及其相邻左/右视角图像（如 CAM_BACK_LEFT 与 CAM_BACK, CAM_BACK 与 CAM_BACK_RIGHT）做 sift 特征点匹配，预测深度，并三维重建。

$$p_t^{i \rightarrow j} = K^j (T^j)^{-1} T^i D_t^i (K^i)^{-1} p_t^i$$

同一时刻相邻视角的特征点投影公式如上：

$pt(i)$ 是第 i 个相机的二维特征点集合， K_i 和 K_j 是第 i 和第 j 个相机的内参， T_i 和 T_j 是第 i 和第 j 个相机的外参， $D_t(i)$ 是第 i 个相机的绝对深度。首先由内参将第 i 个相机的二维特征点投影到其相机坐标系中的三维点，再由外参将第 i 个相机坐标系的三维点转换到世界坐标系；接下来用同样的方法将世界坐标系的三维特征点投影到第 j 个相机的图像坐标系上。

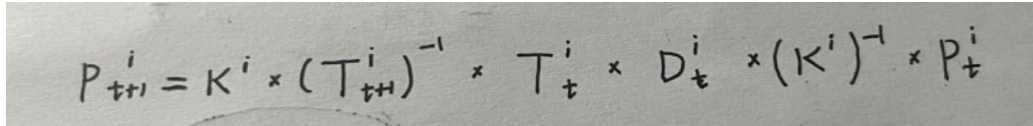
由于 SfM 只是从每张图像与其相邻视角的图像的交叠视野中，寻找特征点做匹配，因此得到的深度图是稀疏的（即深度图 99% 以上的像素点，是没有深度值的，或深度值为 0）。我们用这张稀疏深度图作为深度图真值，来对模型做预训练。

SfM 处理的是同一时刻不同视角的图像，而不是之后正式训练（即第二阶段）的同一视角不同时刻信息，因此在这一阶段不使用预测自运动状态(ego-motion)的运动姿态网络(pose network)

二 . 正式训练

正式训练时，网络已经对物体在真实世界的尺度有大体的认识。与 SfM 不同，正式训练通过运动姿态网络，预测得到每一时刻每个视角的外参。（即 SfM 是同一时刻相邻视角，正式训练是同一视角相邻时刻）

通过测试代码可知，这里的外参 T 是针孔相机相对于初始 $t=0$ 时刻的姿态（平移和旋转），而不是相对于车顶激光雷达的。因此，外参 T 在时刻变化着，需要不断重新计算以更新。


$$P_{t+1}^i = K^i \times (T_{t+1}^i)^{-1} \times T_t^i \times D_t^i \times (K^i)^{-1} \times P_t^i$$

上图公式描述的是单一视角从第 t 时刻到第 $t+1$ 时刻的特征点投影过程，其中， t 时刻称为 source 帧, $t+1$ 时刻称为 target 帧，这种 (source, target) 的情形，同样还适用于 $(t-1, t)$ 的情况。

多视角联合在估计自运动时再一次起到了作用：首先估计多视角联动的自运动，随后再投影到每个单独的镜头上。

$$P_{t \rightarrow s}^i = (T^i)^{-1} P_{t \rightarrow s} T^i$$

三．损失函数与掩码

损失函数主要由两部分组成：一个是 RGB 三维重建损失 reprojection loss，它是逐像素点计算 pred 与 target 的 RGB 图像之间的差异；另一个是单通道深度图的平滑性损失 smoothness loss，它用来惩罚深度图中的不平滑区域（真实深度图应该是相对平滑的）。

由于 SfM 生成的是稀疏深度图，因此在计算 loss 时应只关注那些稀疏像素点，方法是将其他绝大多数没有深度的像素点过滤掉，因此 mask 很重要。论文在代码中提供了三种 mask 的可选项，它们分别是 automasking, fix_mask 和 predictive_mask，其中 predictive_mask 是通过网络预测出的掩码，它的网络结构与 depth_encoder 相同，具体实现中，为每个视角的图像预测一个单独的掩码；fix_mask 则是论文作者期望在输入图像上加上的特定掩码，predictive_mask 和 fix_mask 只有定义，并未在代码中实际实现。

论文复现结果请见第二部分：“基于 SurroundDepth 和 nuscnets 数据集的 supervised learning 改进方法”，可视化视频请见附录中的视频 “nusc_sfm_0.mp4”，视频的具体生成方法与含义请见第二部分。