【深度估计】单目深度估计

文章目录

什么是 深度估计 ？

深度估计，就是获取图像中的场景里的每个点到相机的距离信息，这种距离信息组成的图称之为 深度图 – Depth map

什么是视差

两张图像中相同的物体的像素坐标不同；
较近的物体的像素坐标差异较大，较远的物体的差异较小；
同一个 世界坐标系 下的点在不同图像中的像素坐标差异，就是视差；
不同图像之间的视差，通过相机参数、两个拍摄点之间的位置信息即可换算出物体和拍摄点之间的距离；

深度估计与三维重建

1、获取深度图以及尺度信息
深度图是三维重建的基础
可以通过激光/双目/相机姿态获取尺度
2、将像素坐标转换到世界坐标
通过内参矩阵、外参矩阵以及尺度关系，得到基于世界坐标系下的点云信息
3、三维重建
得到点云后，再将图像的纹理信息贴到点云上，完成三维重建

单目深度估计研究历程

单目深度估计方法

传统方法

基于线索

从图像本身的特征和线索计算图像的深度值。

常用的单日深度线索有：线性透视、聚焦/散焦、大气散射、阴影、纹理、遮挡、相对高度和运动线索。

线性透视

通过检测平行线，识别这些线的会聚点(消失点)来进行深度估计
当距离眼睛更远时，固定尺寸的物体将产生较小的视角
根据消失线和消失点的位置对深度进行适当的分配

聚焦/散焦度

在凸面镜所成的像中，物体只有处在离镜头特定的距离才能够被聚焦，在其他位置都会产生不同程度的模糊现象，模糊程度与其所处的距离有关。

例子：基于聚焦信息构造高阶统计量图，区分出图像中的前景区域和背景区域并对这两个区域进行深度分配。

天气散射

当光线通过大气层传播时，空气中的灰尘微粒对光线具有散射和吸收作用，远处物体相对于近处物体亮度、对比度和色彩饱和度较低，看起来不太清晰。

根据大气散射现象，大脑可以判断不同对比度的物体具有不同的深度。

例子: 通过在输入图像上添加雾面来模拟雾图像，并通过去雾算法中的透射估计方法估计深度图。

阴影

图像中物体表面阴影的变化可以反映物体的形状信息。

SFS(Shape from shading 阴影恢复形状): 利用图像的亮度和表面几何之间的关系，从灰度图像中恢复出物体的三维形状当物体表面的颜色和纹理不属于同一分布的时候，该方法就会失效。

纹理

根据表面纹理标记的提示来估计表面的形状。
距离一个物体越近时，越能清楚地看到物体表面的纹理细节，对于距离较远的物体看不清。
通常仅限于特定类型的图像。

遮挡

当一个物体遮挡住另一个物体时，它比被遮挡的物体距离观看者更近一般认为轮廓线连续平滑的物体是遮挡物体，即距离观察者更近。

例子：通过对遮挡的明确推理，恢复了场景中独立结构的深度排序。

高度

靠近图像底部的物体通常比图片顶部的物体更近，主要包含在户外和景观场景中要提取出这个深度线索，通常要识别出水平线，将图像分成从左边界到右边界的条纹。

例子：应用线追踪算法来恢复最优分割线，并进一步采用深度优化方法来提高最终深度图的质量。

运动线索

基于物体自身运动

利用运动视差近大远小的原理，通过对视频序列的前后帧进行点匹配求得运动视差·只适用于摄像机处于静止的情形，没有运动物体时失效

基于摄像机的运动

运动恢复结构(Structure From Motion ,SFM):假定场景静止不变，仅存在摄像机的运动SFM 可以从图像序列中恢复出摄像机的外参和场景的深度信息
1.首先对相机标定。
2.提取图像特征，并计算相邻图像匹配的特征点。
3.根据对极几何得到相机位姿以及深度信息。
缺点
1.要求必须存在相机的运动，运动幅度不能较大。
2.当场景中存在运动物体时，对精度影响很大；速度相对较慢。
3.依赖相邻图像间的特征点匹配，不适用图像纹理较少或相机的运动幅度大的场景。

基于机器学习

将大量训练图像集和对应的深度图输入定义好的模型中，进行有监督的学习。
分为参数学习方法与非参数学习方法。

参数学习方法

参数学习方法是指能量函数中含有未知参数的方法，训练的过程是对这些参数的求解

开创性工作

2005年，斯坦福大学的Saxena等人利用**马尔科夫随机场(Markov RandomField，MRF)**学习输入图像特征与输出深度之间的映射关系。

利用图像中多尺度的纹理、模糊等深度线索，分别构建了**高斯和拉普拉斯MRF**。

对每个分割图像块的深度进行了建模，同时建立相邻块之间的深度关系。

改进

2007年，在最大化后验概率框架下，以超像素为单元，利用MRF 拟合特征与深度、不同尺度的深度之间的关系，进而实现对深度的估计。

> (超像素:把一些具有相似特性的像素“聚合”起来，形成一个更具有代表性的大“元素”)

加入语义信息

通过**引入场景中的附加信息**，如语义假设和重复纹理等，能有效提高深度估计的精度。

2010年，Liu 等人对整个图像的不同区域按照语义标签进行分类。

采用更简单的特征向量作为监督学习的输入，充分利用不同类别之间的深度信息和几何约束。

将语义信息及对应的深度约束结合，构建MRF模型，优化模型得到场景的深度信息。

MRF通常很难进行精确地学习和推理，大多都采用近似计算，导致预测深度的准确率不高，且效率低。

条件随机场 (Conditional Random Field,CRF)

Cheng 等人首先利用**遮挡**和**消失点**这两种深度线索获取深度梯度图，构建基于像素的条件随机场。

Zhuo等人提出对深度图的分层表达进行建模，对超像素、区域和布局的不同层融合推理。

J等人研究了超像素标记和深度估计之间的内在关系，提出**弹性条件随机场模型Elastic Conditional Random Field，ECRF)**，利用它们的相互关联来加强彼此。

上述方法需假设RGB图像与深度之间的关系满足某种参数模型，而假设模型难以模拟真实世界的映射关系，预测精度有限

非参数学习方法

非参数学习方法，使用现有的数据集进行相似性检索推测深度。

一种数据驱动算法。

给定一幅测试图像，通过融合RGBD数据库中相似图像的深度得到。

第一个里程的工作

Konrad 等人提出采用**最近邻搜索(k Nearest Neighbor，kNN)**。

从RGBD训练库中选出与测试图像最相似的幅候选图像。

再将这 K 幅候选图像对应的深度图进行中值融合得到测试图像的深度。

进一步完善

Karsch等人采用变形步骤，将候选图像和深度与测试图像对齐，构建了【融合变形后的K幅候选深度图的】 能量最小化方程。

基于多帧

利用视频中时间信息来获得时间上一致的深度估计。

Liu等人将单目深度估计视为离散-连续最优化问题。

通过非参数学习方式在数据库中检索相似的深度图，并利用遮挡信息构建目标函数进行深度推理。

其他非参数学习方法

Henera等人使用基于局部二进制模式的特征来估计相似的图像。采用自适应的方法进行融合得到最终深度。

在此基础上他们又提出了基于聚类的深度提取学习算法。

该方法首先根据结构的相似度将 RGBD 数据库进行聚类处理，分割成数个集合。对于给定的输入图像，先找到最相似的图像集计算出先验的深度图，之后采用基于分割的导向滤波对先验深度进行优化。

优点：非参数化方法不需要设计参数化的模型，同时也没有引入太多的场景假设。

缺点：当数据库中不存在与测试图像相似的图像时，很难恢复理想的深度图；依赖于图像检索，计算量大、耗时高，难以实际应用。

深度学习　方法

基于有监督的深度学习方法

基于有监督学习的的单目深度估计方法，在模型训练时需要依赖真实深度依赖庞大的数据进行网络模型的训练，数据集一般包括单目图像和对应的深度真值。

基于有监督学习的单目深度估计方法中，网络模型的训练需要依赖真实深度值。真实深度值的获取成本高昂，且范围有限，需要精密的深度测量设备和移动平台采集的原始深度标签通常是稀疏点，不能与原图很好的匹配。

首次应用

2014年，Eigen等人使用Deep CNN估计单幅图像的深度，两个分支以RGB图片作为输入，第一个分支网络粗略预测整张图像的全局信息，第二个分支网络细化预测图像的局部信息原始图片输入粗网络后，得到全局尺度下场景深度的粗略估计将粗网络的输出传递给细网络，进行局部优化，添加细节信息先训练Coarse网络，再固定Coarse网络的训练参数，去训练Fine网络

一种全局+局部的策略，Coarse网络预测整体趋势，Fine网络局部调优。

改进：多尺度网络

2015年，Eigen等人基于上述工作，提出了一个统一的多尺度网络框架。

使用了更深的基础网络VGG，利用第3个细尺度的网络进一步增添细节信息，提高分辨率，scale1网络对整张图片做粗略估计，scale2和scale3 网络对全局预测进行细节优化，将scale1网络的多通道特征图输入 scale2 网络，联合训练前面两个尺度的网络，简化训练过程，提高网络性能。

分别用于深度预测，表面法向量估计和语义分割3个任务，将同一框架独立应用于不同任务，使用不同的数据集训练。

深度卷积神经场

Liu等人(2015)将深度卷积神经网络　与连续条件随机场结合，提出深度卷积神经场；

使用深度结构化的学习策略，学习连续CRF的一元势能项和成对势能项；

通过解析地求解函数的积分,可以精确地求解似然概率优化问题。

Li等人(2015)提出多尺度深度估计方法，用深度神经网络　对超像素尺度的深度进行回归；

再用多层条件随机场后处理，结合超像素尺度与像素尺度的深度进行优化；

多尺度图片作为输入，有利于学习全局的深度信息。

深度残差网络

Laina 等人(2016)提出一种基于残差学习的全卷积网络(FCN)架构，去掉全连接层，减少参数，不限制图像输入尺寸。

整个网络可以看做是一个encoder-decoder的过程，使用了预训练的ResNet50，网络结构更深。

为了提高输出分辨率同时优化效率，提出一种新的上采样方法。

考虑到深度的数值分布特性，引入逆Huber Loss作为优化函数。

利用分类思想

考虑到场景由远及近的特性，可以利用分类的思想。

Cao等人(2018)将深度估计问题看作像素级的分类问题。

离散化：将深度值投影到对数空间，按照深度范围离散化为类别标签。

训练：深度残差网络预测每个像素对应的类别，损失函数包含信息增益的多项逻辑函数（对离真值越远惩罚越大，网络更加关注难样本）。

后处理：分类可以得出概率分布，便于条件随机场作为后处理优化细节。

基于无监督的深度学习方法

基于有监督学习的单目深度估计方法中，网络模型的训练需要依赖真实深度值。真实深度值的获取成本高昂，且范围有限，需要精密的深度测量设备和移动平台采集的原始深度标签通常是稀疏点，不能与原图很好的匹配。

无监督学习的方法不依赖深度真值，是单目深度估计研究中的热点。

相对于传统算法和有监督学习算法，无监督学习方法在网络训练时只依赖多帧图像，不需要深度真值具有数据集易获得、结果准确率高和易于应用等优点。

根据图像对之间的几何关系重建出对应的图像，通过图像重建损失监督训练。

利用立体视图

Garg等人(2016)提出利用立体图像对实现无监督单目深度估计；

利用左右立体图像对，用预测的深度图重构左图，计算重构损失；

训练时需要左右图像对，预测时只需要一张图；

Godard 等人(2017)对上述方法进一步改进: Monodepth；

利用左右视图的一致性实现无监督的深度预测；

利用对极几何约束生成视差图，再利用左右视差一致性优化性能，提升鲁棒性。

利用相对关系

Zoran等人(2015)关注相对深度关系，利用图像中点对之间的相对关系推断深度信息。（需要少量相对远近的标签，算是弱监督）

网络输出点对之间的相对关系，再利用数值优化方法将稀疏的输出稠密化为最终结果。

优点：比数值回归更加简单；人们能够很容易判断相对关系，训练数据集获取成本低相对关系不受数据的单应变换影响，系统更加鲁棒

整体框架由3部分组成:
第1部分从图像中选择点对。
第2部分估计每一个点对的相对关系，提取相关信息并做三分类。
第3部分将点对之间的相对关系扩展至全局，得到稠密输出。

Chen 等人(2016)利用相对深度关系构造损失函数通过多尺度的神经网络直接预测像素级的深度。

此损失函数的设计，让网络能够利用相对深度关系作为标签，深度值作为网络的输出结果，将相对深度关系与连续深度值联系了起来。

利用视频序列

SFMLearner
Monodepth2
Featdepth

应对动态障碍物

上述方法都基于静态场景假设，如果场景中出现了动态目标，动态目标在两帧中的变化就会很小，可能将近处的物体误判为远处的物体（因为远处的物体误差小）。

Struct2Depth

Depth in the Wild

Depth and motion learning

数据集

KITTI

NY U

CityScapes

论文推荐

CVPR 2021