多目深度估计（Multi-view stereo）

1.MVSNet: Depth inference for unstructured multi-view stereo(2018)

论文链接

读的第一篇深度估计的论文，结果一开始被单应性拒之门外，先恶补了许多图形学基础知识，又读了一些博客和pytorch代码解读，如今读来才感觉把整个pipeline搞懂了。本文提出了MVSNet，基于深度学习的MVS方法

- Multi-view stereo输入一张参考图和多张源图片来预测参考图的pixel-wise深度，训练时每个视图分别作为参考图
- 网络结构：用一个shared CNN来为每张图片(3*H*W)提取特征图(H/4*W/4*F)；对每个特征图在一定深度范围内均匀取样D个，并分别按照对应深度的单应矩阵将source image的特征图用单应性变换统一到参考图片的平面，沿深度拼起来得到V_i，H/4*W/4*D*F；利用方差将所有V_i计算统一得到一个cost volume C，尺寸与V_i一致；将C经过3D卷积的类U-Net网络，进行正则化，最终经过一个1*1卷积，得$H/4*W/4*D'，沿深度做softmax，得到概率图，每个像素对应一个D维的概率向量（对应当前深度的概率）；沿深度做soft argmin，其实就是对每个像素的概率向量取深度的期望，得到intial depth map；将ref image resize成H/4*W/4*3,与intial depth map(H/4*W/4) concatenation起来，经过几个卷积层与intial depth map相加，获得refine后的depth map

> 一个没能解决的疑惑是，最终的深度图是H/4*W/4，不知道咋恢复到原尺寸，或者不用恢复？从R-MVSNet论文中看，最终深度图确实下采样了四倍

- 损失函数：对intial depth map和refine后的depth map，分别累加有效像素的预测深度与GT的L1范数
- 概率图：计算4个最近邻深度的概率求和，用来估计深度预测的质量，得到概率图
- 后处理：深度图过滤。过滤去异常值
  - photometric consistency：衡量估计质量，过滤去p<0.8的像素点
  - geometric constraint：将参考图中的像素点p1投影到一个source图中p_i，再将p_i重投影回参考图p_1'，若p_1和p_1'的坐标值差值和对应深度差值都在某一阈值下，则称为两视图连续，本实验中每个像素点至少三视图连续

- 后处理：深度图融合。N个视图分别预测出深度图，将每个像素重投影得到的每个深度图的深度取均值，作为最终深度估计。
- 论文中给出的单应性矩阵的公式有误，具体见https://zhuanlan.zhihu.com/p/363830541（这里边应该也给反了，给成从ref到source了）

2.R-MVSNet:Recurrent mvsnet for high-resolution multi-view stereo depth inference(2019)

论文链接

在MVSNet的基础上，加入了GRU，提出R-MVSNet，创新点有限

- 背景：MVSNet效果很好，但是在使用3D卷积对cost volume正则化时，太费内存了，也因此难以应用到高分辨率的图像。R-MVSNet用GRU替换3D conv来对cost volume正则化，有效减小了内存消耗

- 网络：大致与MVSNet类似，在获得cost volume后，沿深度方向接入三层GRU（卷积变体），尽管感受野变为当前深度及之前的特征，但效果类似，最后一层的输出通道数为1，经过softmax获得概率volume，与GT计算交叉熵损失，作为分类任务训练

- 为了增大深度估计的范围，R-MVSNet在采样深度时使用了inverse depth，而不像MVSNet在深度范围内均匀采样。inverse depth是按深度的倒数取样，也许是对深度的倒数均匀采样（论文中没有明确说明）。也因此，不能用soft argmin来回归获得预测深度（深度采样不均匀），而采用了一种分类，并refine细化的方法

- 后处理：因为按照分类任务训练，预测时相当于argmax取深度，无法获得亚像素级深度估计，而且预测的深度图不连续，会有阶梯效应。因此引入一个 Variational Depth Map Refinement，类似于插值，使深度图变得smooth。
  在refinement种引入了一个reprojection error，包括两部分：

  - photo-metric error:将source image I_i根据ref的深度图D_1（感觉这里存疑，投影方向不对吧）投影到I_1,用zero-mean normalized cross-correlation衡量二者的error
  - 正则化项：将每一像素与其相邻像素计算并累加bilateral squared depth difference，这一项会使深度图smooth

  论文中提到，会迭代地使这个error最小化，具体怎么做不清楚

3.Point-based multi-view stereo network(2019)

论文链接

本文提出基于点云的MVS方法Point-MVSNet

- 背景：MVSNet中使用3D卷积正则化cost volume，利用了3d特征但内存消耗太大，Point-MVSNet可以在避免3D conv低效的同时利用3D几何特征
- 思路：先基于MVSNet方法预测粗略的深度图，反投影成3d点云，结合2D特征后应用Pointflow预测点与GT的残差，用残差修正并细化点云，再迭代这一过程
- coarse depth prediction:与MVSNet相比，本文中的下采样倍数由4变为8，通道数也大大下降，因此3D卷积的内存消耗大大下降
- 2D-3D特征融合：

    - 在CNN提取特征图时，每张图片提取三个尺寸的特征金字塔F_i。先将不同图片的特征图投影到同一平面（根据相机内参矩阵和外参），再对分别每个尺寸的所有图片的特征图取方差运算来统一成C_i，作为2D特征。对于粗糙深度图生成的3D点云，对每个点将世界坐标X_p和投影到C_i上对应的特征concatenation起来作为点的特征
    - 由此，融合了多尺度的2D特征，和3D几何特征，获得特征加强点云
    - 并且，每次迭代更新点云后，提取的2D特征会有所不同，实现dynamic feature fetching
- Pointflow:对于3D点云的每个点，沿投影方向以s为间隔生成2m个假设点（即深度间隔为s），通过对每个点做边卷积，再经过MLP和softmax获得每个假设点所在深度的概率。最后，对每个假设点的概率乘ks（与非假设点的间距）并累加，获得间距的期望，从而获得残差深度预测。与原深度图相加可以获得细化的深度图，再迭代这一过程，让点云中的点"flow"向GT。
- 每次迭代会对深度图进行上采样（最近邻），以获得更高分辨率的深度图，并减小间隔s，以捕捉更细的特征。本文只迭代两次

4.cascade MVSNet:Cascade cost volume for high-resolution multi-view stereo and stereo matching(2019)

论文链接

在MVSNet的基础上，提出了Cascade cost volume，并应用FPN的思想，优化内存和时间效率的同时，提高了效果

- 思路：使用FPN提取每张输入图片的不同尺寸的特征图（3个），分三个阶段。从最top层开始（分辨率最低的，1/16），构建cost volume,输入MVSNet，回归得到深度图；将深度图上采样得到与下一阶段特征图一致的尺寸，以上一阶段深度图为中心，确定深度范围/间隔，对深度进行采样（即每个像素点p的深度假设为$d_p+\delta$，从而可以为曲面假设）进行单应性变换，得到cost volume，再输入MVSNet，重复得到第三阶段输出的深度图（分辨率与原图一致）
- 每一阶段比前一阶段的深度范围缩短，且深度间隔变小，分辨率增大。因此，尽管第一stage和经典MVSNet的深度范围/间隔差不多，但分辨率低所以内存消耗更小。后边的stage深度范围大大下降，从而内存消耗也小
- 损失函数为每个阶段深度图损失的加权和

5.P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo(2019)

论文链接

本文提出了一种新的建立cost volume(MCV,matching confidence volume)的方法，P-mvsnet达到sota

- 背景：过去的cost volumn不考虑参考图像素，且为pixel-wise（对噪声鲁棒性不好），cost volumn应为各向异的，但原本简单的方差求法为各向同的。本文提出了一种新的计算cost volumn的方法，patch-wise，且各向异，并应用了各向同的3D U-Net进行正则化
- 网络：
    - 先将图片进行下采样4倍的特征提取；
    - 单应性变换后，计算每个src与ref的MSE，将结果取负后取指数得到pixel-wise的MCV，每个像素预测的是当前深度假设的置信度，输入一个patch-wise matching confidence aggregation module，聚合每个像素点所在patch的特征和相邻深度对应patch的特征，计算置信度，聚合的过程是可学习的（卷积），因此提高了鲁棒性；
    - 将patch-wise的MCV输入3D U-Net，其中包含各向异的3D卷积层（如1*3*3、7*1*1），得到LPV(latent probability volumn)，将LPV softmax后获得PV(probability volumn),深度回归得到深度

预测值（期望）；
    - 最初src的特征图经过一个解码器上采样（反卷积）两倍，再与经过上采样的LPV concatenation，经过卷积层refine后可以得到更高分辨率的深度图
    - 损失函数为两张深度图预测损失的加权和
- 后处理（点云重建）
    - Depth-confindence:除去明显不可信的预测，用PV衡量置信度，抛弃两个PV(argmax)之和小于0.5的预测
    - Depth-consistency:先将ref上的像素点p根据预测深度投影到一个src上p'，再将p'反投影回来，计算与p的深度差和坐标差来表示一致性。
- 在Depth-consistency中，本文讨论了一个很有意思的地方，也是我之前疑惑的一个点，如何将p'反投影回去。这涉及两个问题，一是src图没有深度图，取不到p'的深度，二是p'不一定刚好的像素点上，有深度图也没用。本文提出了三个方法：（提到的深度均为ref的深度图）
    - nearest depth：取离p'最近的像素点和其深度反投影（太朴素了）
    - bilinear depth：取p'的深度为临近四个像素点深度的双线性插值，将p'反投影回去（文中提到，当GT相机参数已知时用这个，否则用的下边这个）
    - depth-consisten first depth:取p'邻近的四个像素点中深度与p最近的，将其按对应深度反投影回去

6.CVP-MVSNet:Cost volume pyramid based depth inference for multi-view stereo(2020)

论文链接

引入image金字塔和cost volum金字塔，用coarse-to-fine方法，提出CVP-MVSNet(cost volum pyramid)，和Cascade-MVSNet挺像的

- 网络：取L个level的图片集合构成图片金字塔；分别对每个level提取特征图；从最top level开始（coarsest），单应性变换后根据方差构建cost volumn，经过3D卷积后回归得到coarse深度图；将上一阶段深度图上采样后，以此为中心确定新的深度采样平面，根据当前level的特征图构建cost volumn，之后和Cascade-MVSNet类似预测的是残差深度图，修正后输入下一阶段
- 深度采样：除第一阶段外，深度采样的范围、间隔都由上一阶段深度图确定。深度间隔通过计算0.5像素内深度差的均值，深度范围为将ref上的点投影到src图，并将对称的假设点也投影到src图（由对极约束，必在极线上），当src图极线上的投影范围恰好两像素，此时的边界假设点即深度范围
- 损失为每阶段深度图损失的加权和

7.Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement(2020)

论文链接

本文兼顾MVS任务的效率和效果，进行由粗到细、由稀疏到稠密的预测，提出Fast-mvsnet

- 思路：先经过2D CNN提取特征，构建cost volumn，经3D卷积正则化后，获得高分辨率、稀疏的深度图；将深度图进行传播，来使其变稠密，先对对稀疏深度图按最近邻加密，再以ref图为输入用CNN为每个点预测一个k^2的权重（输出尺寸与深度图一致），再求最近邻k^2个点深度的加权和；将稠密的深度图进行高斯-牛顿细化获得最终深度图，这个过程综合了初始特征图，并且没有需要学习的参数
- 稀疏特征图：低分辨率的深度图没有细节，高分辨率深度图计算成本太高，因此本文先计算稀疏高分辨率深度图，再细化，既节省成本又获得高分辨率。具体实现为，按低分辨率构建cost volumn并预测深度，再稀疏化，后经深度图传播获得稠密深度图

8.UCS-Net:Deep stereo using adaptive thin volume representation with uncertainty awareness(2020)

论文链接

和cascade MVSNet非常像，特征提取的CNN用的U-Net，第二/三阶段的深度采样范围依赖于前一阶段深度图的方差，这个适应特点在文中被称为ATV(adaptive thin volumn)。通过逐阶段提高分辨率（上采样）和细化深度采样，来refine深度图

9.Patchmatchnet: Learned multi-view patchmatch stereo(2021)

论文链接

本文将计算机视觉中的patchmatch方法应用到MVS任务，利用可学习的patchmatch模块coarse-to-fine，在效果与sota差不多的同时大大降低了时间和内存消耗

- 思路：先利用FPN提取多尺度特征图，将分辨率最低的特征图输入patchmatch模块，获得深度图，上采样后指导下一阶段的patchmatch，如此级联。最后一个阶段的深度图利用ref image细化，得到原始尺寸的深度图
- learning-based patchmatch:分三步，初始化，传播，评价，在传播+评价迭代多次直到收敛
    - 初始化：和R-MVSNet类似，用inverse depth采样F_f个深度假设平面
    - local perturbation:第二阶段开始，以上一阶段的深度图为中心，采样N_k个深度假设，深度范围也会细化
    - adaptive propagation（核心）:其实也是采样深度假设，对特征图的每个点，取邻近的K_p个点（使用fixed偏移，也许是网格），将其深度（上一阶段特征图）作为深度假设。本文进一步加入了自适应特点，希望采样的这K_p个点可以在同一平面（近似），因此用一个2D CNN为ref image的每个像素点学习了一个K_p的偏移，加到fixed采样的K_p个点的坐标上，作为最终采样的点，用其深度作为深度假设。我们希望这个自适应的偏移可以令fixed采样点修正到与待测点位于同一平面的位置，由此得到更好的深度假设。
    - Matching Cost Computation:利用通过上面两个方法得到的深度假设将特征图投影到ref平面，计算match cost。先分别对每张图的特征图（W*H*D*C）操作，将C个维度分成G组进行相似度计算，得到S(W*H*D*G)，利用一个3D卷积计算置信度P(H*W*D)，对P取max，得到pixel-wise view weights w(H*W)，w仅计算一次，后续通过上采样即可。用w作为S的加权计算所有图片S的均值$\bar{S}$(W*H*D*G)，利用3D卷积获得成本C(H*W*D)
    - Adaptive Spatial Cost Aggregation:对每个点采样邻近的K_e个点（网格），再用CNN预测一个K_e维的偏移来修正以作为最终采样点，将不同采样点的成本根据特征相似度和深度相似度加权求均值，得到聚合空间成本
    - 对聚合空间成本使用softmax获得概率体，沿深度求期望得到深度图，再迭代传播+评价这个过程（文中三个阶段的迭代次数分别为221）

10.TransMVSNet: Global Context-aware Multi-view Stereo Network with Transformers(2022)

论文链接

第一篇将transformer应用到MVS任务，提出TransMVSNet，网络结构（论文中的图特别明确）：

- 利用FPN提取不同尺寸的特征图

- 将不同尺寸的特征图分别输入ARF(adaptive receptive field)，来缓解FPN和transformer之间感受野的gap，具体为用deformable convolution扩大感受野
- 将top层的特征图（分辨率最低）输入FMT(feature matching transformer)，加上位置编码后flatten，输入N_a个级联的transformer块，在每个块内先计算每张图片特征图的self-attention，再计算ref与每个src的cross-attention，这里改变的是src的值（src作为query），为了保证不同src查询的ref值不变
- 为了节省计算成本，仅top层特征图会经过FMT，之后通过transformed feature pathway将低分辨率的特征图（已经过FMT）上采样后和高分辨率的特征图（经过ARF）加起来。
- 分别对每个尺寸的特征图进行深度假设采样和单应性变换，统一到ref平面，利用pair-wise feature correlation分别计算src和ref的correlation volumn，再经过加权和（权重为在文中有说）计算聚合correlation volumn
- 将聚合correlation volumn经过3D卷积层得到probability volumn，使用argmax获得深度预测，使用focal loss
- 低分辨率的深度图在上采样后和下一阶段的特征图结合，实现coarse-to-fine预测深度图

> 有个有意思的观点，本文认为MVS本质是一对多的匹配问题，因为当ref,src的相机确定，对任意可能的深度，由对极约束ref上的点p对应的点必在src的极线上，相当于p与极线上的候选点的匹配

图形学拾遗

- 刚体运动（2D/3D）：旋转/平移/刚体/缩放/仿射/透视变换矩阵，齐次坐标，旋转向量/欧拉角/四元数
- 相机模型：针孔相机模型，世界/相机/像素坐标系，内参矩阵/外参数，畸变
- 2D-3D对极几何：对极几何约束（极线等），本质矩阵（八点法求解），单应矩阵，单应性变换