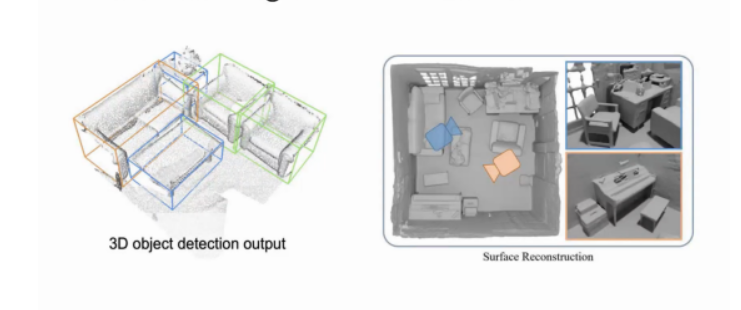
****

**韩晓光（香港中文大学深圳）：**这个图的左边是比较常见的室内场景点云，围绕这个有很多做分割以及3D检测等理解类的任务。右边是关于重建的示意图，例如近年的像Neural RGB-D Reconstruction以及NeuralRecon等工作。针对目前这些重建工作来说，它们并不会去进行理解，不会去关心场景里面是什么物体。我个人觉得这两件事情应该是需要结合去考虑的。比如说我们最近这两年做的一些工作，例如Total3DUnderstanding (CVPR 2020) 是从单视角图像进行全场景的重建以及RfD-Net(CVPR2021)是从点云中进行重建，这两个工作都是先对场景进行理解，检测出物体实例，再对每个实例进行重建。并且，我们也发现理解可以帮助重建，而重建也可以反过来帮助理解。虽然目前的可视化结果还很不好，但我还是觉得我们应该是要这么去做。最新的我们ECCV 2022的一个工作，叫InstPIFu，也是从单视角图像进行重建，我们基于PIFu的思想，利用Pixel-aligned的特征，对于每个物体实例做了更加精细的重建，比起Total3D以及浙大崔兆鹏老师他们的Im3D（CVPR2021）可以达到更精细的重建结果。这也是我们在这个方向一直想努力的目标，就是提升基于理解的重建精度。

**刘利刚（中国科学技术大学）：**所以原来做得不好是因为空间不够大，还是不好优化？关于这个问题，我是这样理解的：输入图片无论单张还是多张，重建几何是一个欠定问题，即是一个多解问题。只不过我们一直在找所希望的那个解，但是有不同的几何、不同的材质可以生成同样的图片，无论渲染方程是面的还是体的。这个解虽然在那里的，但是不容易找到，于是发明了NeRF这种方法尽可能地去找合理的解。之前NeRF++那篇文章提出了一个radiance ambiguity的概念，几何是球，加上不同的纹理，它也能呈现出相同的图片。

Instant NGP、Mobile NeRF

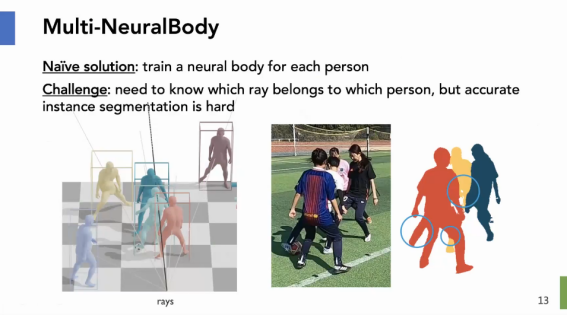
大家读过一篇文章叫Plenoxels就在讨论到底是 MLP的表达重要，还是这个volume rendering重要，它得出的结论就是后者。所以现在的优化其实很强的，它没有用MLP，它就用了一个稀疏的一个体素结构，而且它优化每一个体素里面的球谐函数，它直接就能把这个场景结构恢复出来，所以说volume rendering是最重要的。

本身现在的NeRF rendering它是没有引入先验的，只是说图片直接去往里面比较

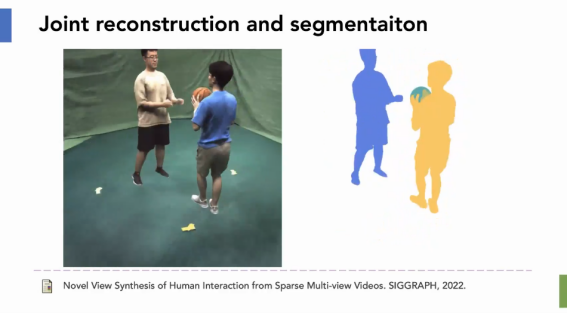
**许威威（浙江大学）：**我介绍它主要是想告诉大家，结合先验我觉得还是挺重要的一个事儿，否则你就是从头弄，我觉得很难加速。如果你没有先验，那就变成从头训练，从头优化就是会影响你的一些效果。第二个就是，你是可以去拿MLP这个基础表达去恢复原来的表达，我是选用了一个别的文章，这个文章建了好多的MLP，它优化的时候刚开始是用volumetric，当几何出来之后它就改用mesh。 这样质量会好很多，会加速。其实我感觉这结果是不错的，能重建出很多细节。我是赞成把NeRF倒回去跟原来CG流程接起来，你不一定说要摆脱原来的流程成为一条独立的流程，这个是两个选择。当然你也可以独立，就看你在不同的应用环境下哪个得到益处多了。但是，如果你想进入原来游戏的市场你就得和原来的CG流程对接。

那么接下来就举几个例子，我们最近的一些工作。第一个就是跟NeRF-in-the-wild很像，所以我们自己叫NeuS-in-the-wild。 NeRF-in-the-wild做的是novel view synthesis，我们这里还是希望从大量的互联网图片重建出几何。其实以前MVS那一套也做过，叫做Building Rome in a Day，但只能出一个点云。对于这种互联网图片，因为它光照变化或者视角变化都比较大，所以直接跑COLMAP很难得到一个干净的几何。但是现在用NeuS这一套的话，就可以看到重建出来的连续性，光滑性是很好的，虽然细节有些丢失，但从互联网图像的角度来说质量已经挺好的。这里能做出来的还有一个原因是我们可以表示光照变化，就像NeRF-in-the-wild。

另外一个例子就是对人体的建模。这个是Neural Body，我们去年的工作，从稀疏的相机拍摄的视频去恢复一个动态的人的NeRF。以前对人的重建基本上还是MVS那一套，因为很难去利用人体的先验，也很难去利用时序的信息，它基本还是当成静态去做的，所以需要大量的视角去拍摄。那么现在用神经表示这一套，首先可以引入人体的一个先验。其次可以利用人的运动的先验，整合整个视频里所有帧的信息。比如Neural Body的做法就是假设每一帧的神经辐射场都是同一组隐编码decode出来的，而且这一组隐编码它贴在一个参数化模型上面，会随着每一帧人的pose不一样，编码的空间位置也会发生变化。那么这样的话，就可以用同一种隐编码去表示视频中所有帧的一个人的NeRF，那么就可以利用视频中所有时刻的观测来做一个更好的重建，这样的话也可以减少每一帧的输入视角数。另外，因为隐编码的空间位置跟参数化模型一样，也相当于用到了人体的一个大概的几何先验。所以，用神经隐式表示这一套可以很好地加一些重建对象的先验进去。



另外就是的今年Siggraph的工作，我们把neural body扩展到了多人，其实也很简单，就是每个人建一个neural body。但是这有一个问题，如果多人去这样做，还是需要分割。因为要训练多个neural body的model，得知道每一帧里面每个像素对应到哪个人。但用一个预训练的分割网络，在这种多人的有遮挡的室外的情况下，其实很容易出错。怎么做呢？这里的一个关键就是，每个人是独立运动的，它跟背景也是独立的。基于运动不一样，我们可以把他们分开。所以这里的关键就是说，我也可以用网络在3d空间里面表示每个点它属于哪一个（部分），然后去同时优化多人的 NeRF，以及每个点它到底属于哪一个人。因为motion不一样，每个点只能被一个人表示，所以这个也是可以一起去优化的。左边是我渲染的结果，右边是Mask，这个不是用预训练网络割出来的，是优化出来的，可以看到mask其实是很精细的，甚至比手标的可能还要精细一点。



所以这个例子我想说吗明，在很多时候，如果我们对场景能做一个这样的重建，其实对我们AI的算法也很有帮助。比如说我们重建很多这样的场景，就相当于免费的得到了很多人体的mask标注。所以我觉得NeRF或者NeuS这一套有一个潜在的很重要的应用，就是给AI提供数据。现在像NeRF这一套，它本身建模是从真实的数据来的，所有最后合成出来的东西会比较真实，这是第一。第二就是它可以基于这个去生产很多新的东西，比如说生成新光照，或者生成新的一些场景的布局等等。第三个如果在重建的过程中还能得到一些free的label，那就更好了，就不用标了，所以现在很多做无人驾驶的在关注这一套，希望能用这个来生成大量的训练数据，或者创造一个高真实感的仿真，来帮助训练AI。

 但我觉得NeRF做到那一点还很长，4k8k的分辨率就是一个很大的瓶颈。光训练都已经需要那么高清，要让NeRF生成的那么高清，现在我觉得暂时是没法达到的。

NeuS

 现在NeRF还有很多缺陷，训练很慢，解码速度也不行，所以有很多工作要做。

逆向的是一方面，在正向上，在图形领域从模拟建模、模拟仿真。它是这样的一种表达方式，不见得叫NeRF，应该是神经网络的隐式表达。