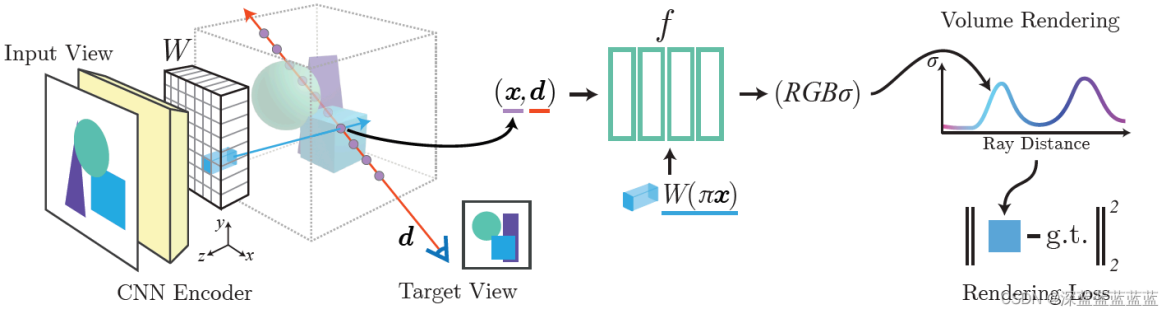
pixelNeRF、

> pixelNeRF: Neural Radiance Fields from One or Few Images(CVPR 2021)

链接：https://arxiv.org/pdf/2012.02190

本文是针对NeRF的优化，NeRF中往往需要同一个场景中非常多的图片才能很好地生成新的视角，而pixelNeRF就提出了一种方式来让模型在仅有几张甚至一张图的情况下也能生成新视角。

文中提出，NeRF之所以无法在一张图基础上生成新视角的原因是没有**先验信息**。因此本文预先通过resnet对输入图片提取了基于每个像素的信息，然后在生成新视角时通过查询对应像素上的信息从而获得先验，辅助NeRF生成新视角的图像。



上图就是在网络仅有**一张图片作为输入**时网络的运行流程，最中间绿色的那个f其实就是NeRF，因此与NeRF唯一的区别其实就是CNN Encoder。

网络运行流程如下：

1.首先给定一张图片，然后使用预训练好的CNN Encoder（ResNet34）提取出图像中每个像素的特征，构建出W，即特征图。

2.然后给出一个想要生成的新视角的相机内参d，按照NeRF中描述的方式往空间中发射一道光线，将交点x投射到输入图所对应的平面上，并从特征图中提取对应的特征。

3.然后将提取出的特征和x，d一起输入NeRF，从而得到最终的颜色和密度信息。

4.最后通过体渲染来的到最终结果，与Ground truth做损失函数来优化整个网络。

上面说的是输入一张图时的做法，但其实这个模型也可以处理输入**多张图**的情况，具体做法就是从每张图中都提取一个特征图，然后将交点x投射到每个平面上，进而从每个特征图中都提取一个特征。最后将每个特征都分别送入NeRF，对结果做聚合再得到最终的颜色核密度。

具体的流程就是上面所说的那样，但有些点需要详细说一下：

首先，Pixel可以在多视图图像的数据集上面进行训练，而不需要任何额外的监督；

其次，PixelNeRF预测输入图像的摄像机坐标系中的NeRF表示，而不是标准坐标系，这是泛化看不见的场景和物体类别的必要条件，因为在有多个物体的场景中，不存在明确的规范坐标系；

第三，它是完全卷积的，这允许它保持图像和输出3D表示之间的空间对齐；

第四，PixelNeRF可以在测试时合并任意数量的输入视图，且不需要任何优化。

第五，在提取特征的时候为了获取局部和全局的信息，作者同时提取了resnet中四个池化层的特征，因此每个像素中都包含了局部和全局的信息

第六，特征输入NeRF的方式是通过类似residual的方式来引入的，并不是直接和x,d合并

一些想法

个人认为这个提feature的思想其实和NLP中对单词做embedding没有区别，只不过NLP中针对的是每个词语，而这里针对的是每个像素。通过大量相似场景的训练来让模型获得一些先验（类似于人的直觉），进而获得想象图片中蕴含的3D信息的能力，从而从feature中预测出不同视角中的信息。

设想一下，如果没有这个预先提取的特征图，直接按照文中描述的方式训练能得到什么样的结果？如果不提取特征图那输入nerf的额外信息就只有对应像素上的一个RGB值，nerf无法知道这个像素周围都有什么东西，因此像素之间无法建立起关联。因此，所谓的先验肯定都是从feature中获得的。

但是，文中的feature是使用预训练的resnet提取的，这提取的feature真的适用于这个任务吗？如果使用一个网络来专门学习特征提取会不会更好？