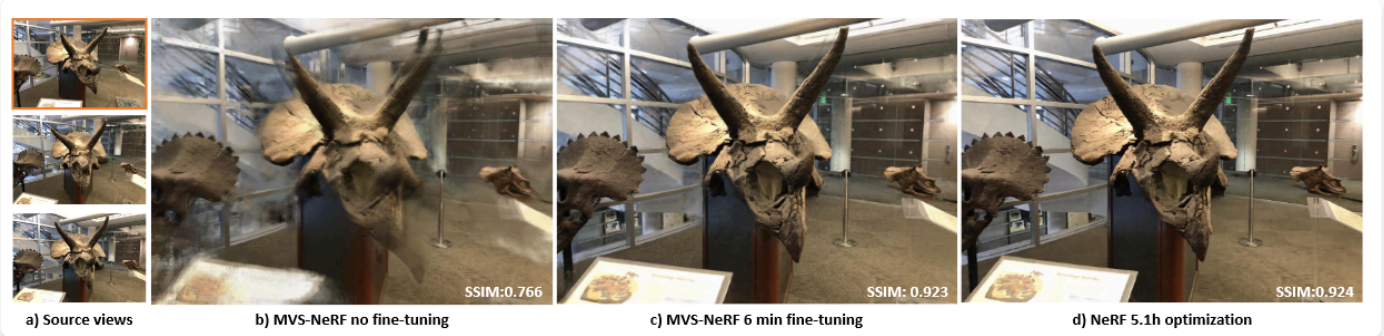
MVSNeRF

论文：《MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo》
地址：https://arxiv.org/abs/2103.15595
年份：ICCV 2021

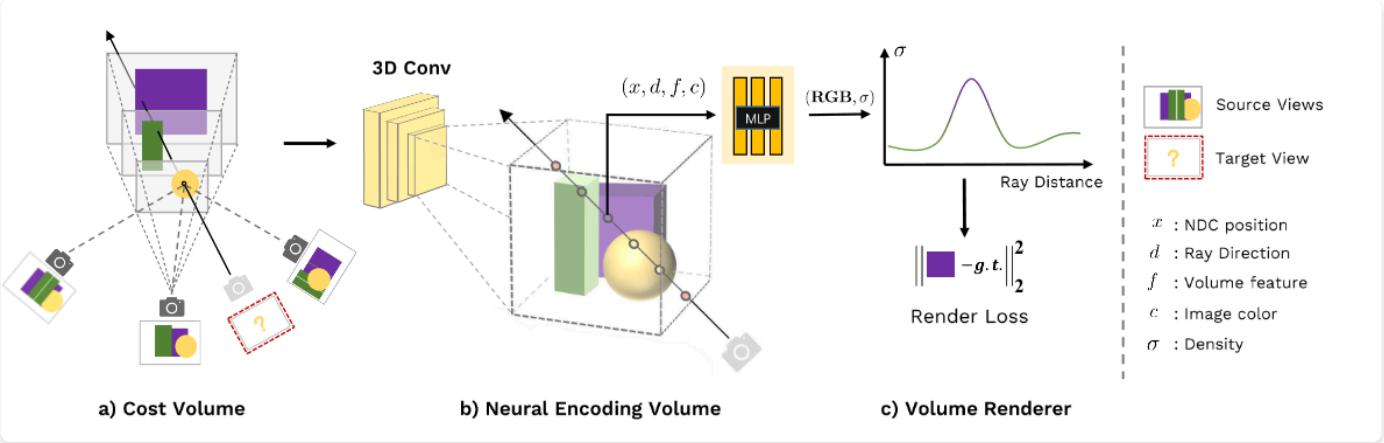
Introduction

任务：(可泛化的) 新视角合成
技术贡献：
(1) 引入 MVS 中的思想，构建一个 feature volume 表示场景，在此基础上实现场景间的泛化。

Method



MVSNeRF 能在一个数据集训练，然后泛化到其他数据集，但如果要在其他数据集上有比较好的效果，还是要进行微调，但微调时间相比重训一个 NeRF 还是非常短的。如上图所示，MVSNeRF 仅需 6 分钟的微调即可达到较好的效果。文章设置训练输入为 3 张图像，微调的输入可以是稠密的图像。



MVSNeRF 的训练 pipeline 前半部分与 MVSNet 基本一致，提取输入图像特征，构建 cost volume。后半部分也比较简单，经过 3D CNN 得到 neural encoding volume，在这个 volume 中进行 ray marching，将点的坐标和视角输入到 MLP 中得到 σ 和 $\mathbf{c}$，渲染得到图像，计算 loss，进行端到端的训练。finetune 过程稍有不同，会在后面提及。

Radiance field reconstruction

cost volume 构建过程与 MVSNet 基本一致，所以从构建 neural encoding volume 开始。

Neural encoding volume

使用一个 3D CNN B 将 cost volume P 转化为 feature volume S:

$$S = B(P)$$

Regressing volume properties

我们认为 feature volume 中包含了场景相关的信息，给定 3D 坐标 x，和视角方向 d，使用一个 MLP A 来得到 σ 和 $\mathbf{c}$。但存在一个问题，在得到 cost volume 时提取了图像特征，特征图的分辨率是原图像的 $1/4$，因此 feature volume 分辨率较低，很难从中恢复出高频信息。文章提出的解决方案是也将原图像对应位置的像素值输入到 MLP 中。将 x 按视角 i 进行投影，得到对应图像的像素值 $I(u_i, v_i)$，按视角拼接成一个 c:

$$\sigma, r = A(x, d, f, c), f = S(x)$$

其中 $f = S(x)$，是在 x 处对 volume 值进行三线性插值的结果。x 也会被转换为 reference view 下的 ndc 坐标，使用 ndc 坐标系相当于对场景进行了归一化，能够消除不同场景尺度的差异，得到更好的泛化性。

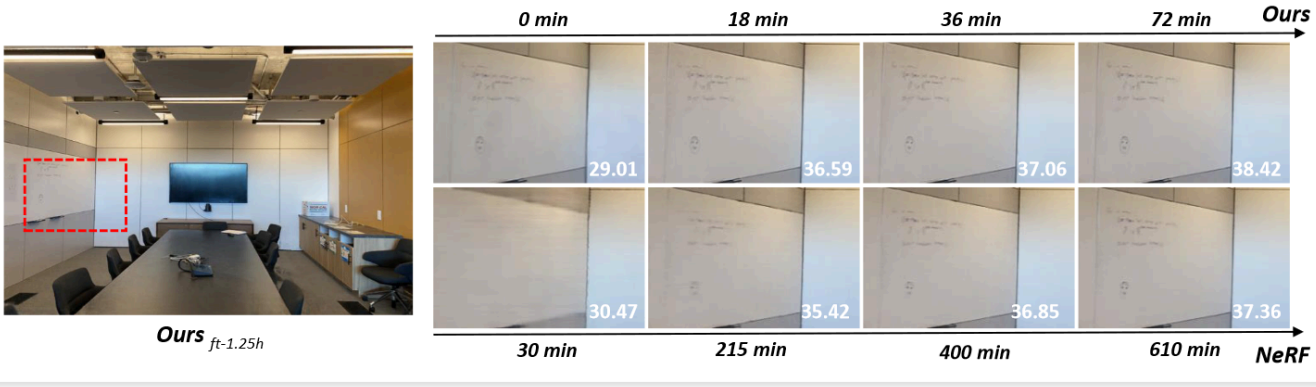
Optimizing the neural encoding volume

先总结一下训练过程中会有哪些东西得到训练：用于提取图像特征的 2D CNN，用于构建 feature volume 的 3D CNN，输出体密度和颜色的 MLP。在 finetune 过程中，我们训练的是 feature volume 和 MLP。

Appending colors

TBD.
可能是用训好的 CNN 先得到输入图像的 feature volume [N, 8, D, h, w]，然后将 volume 中每个 voxel 投影到各个视角下得到像素值 [N, 9, D, h, w]，拼到 feature volume 上作为额外的通道，然后在 finetune 时颜色值会作为 feature vector 也得到训练。

Experiments


0 min
18 min
36 min
72 min
Ours
29.01
36.59
37.06
38.42
30.47
35.42
36.85
37.36
30 min
215 min
400 min
610 min
NeRF

Ours *ft-1.25h*