DVGO笔记

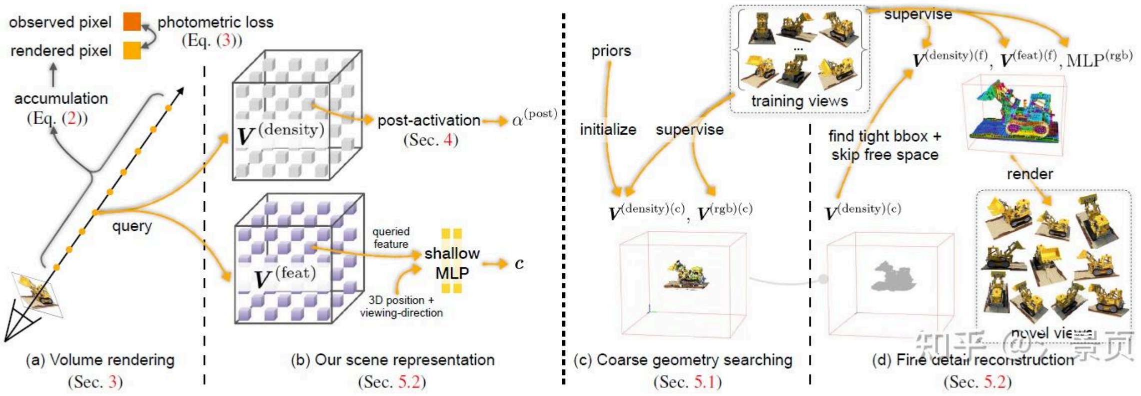
> Direct Voxel Grid Optimization: Super-fast Convergence for Radiance Fields Reconstruction(CVPR 2022 Oral)

原文地址

主要贡献

DVGO 有几点贡献：

1. 用网格存取特征取代了 Encoding（我个人认为和 Instant-NGP 的 Hash Encoding 是一个性质的）。
2. 采用一个网格直接存储了 density 信息，类似于 Mip-NeRF 360 中 Proposal MLP 的作用。
3. 三线性插值后过一个 SoftPlus。
4. 分了两个阶段训练。



> Figure: Approach overview. We first review NeRF in Sec. 3. In Sec. 4, we present a novel post-activated density voxel grid to support sharp surface modeling in lower grid resolutions. In Sec. 5, we show our approach to the reconstruction of radiance field with super-fast convergence, where we first find a coarse geometry in Sec. 5.1 and then reconstruct the fine details and view-dependent effects in Sec. 5.2.

最开始没弄清楚的名词

- 校准图像（calibration images）

说白了就是用于相机校准的图像。这些图像帮助确定相机的内参（intrinsic parameters）和外参（extrinsic parameters），从而确保通过不同角度拍摄的图像可以准确地反映物体的三维结构，消除镜头畸变并精确测量图像中像素点对应的真实空间坐标。

可以看这篇

- cross-scene pretraining

Cross-scene pretraining（跨场景预训练）的核心思想是：在多种不同场景上进行预训练，以提高模型在新场景或未见场景上的泛化能力和性能。

好处如下：

提高泛化能力：通过在不同场景的数据上进行预训练，模型能够学习到跨场景的共性特征，提高在新场景上的表现。

减少过拟合：模型在单一场景上训练时容易过拟合特定场景的特征，而跨场景预训练可以让模型学习到更通用的特征，从而减少过拟合。

提升效率：在某些任务中，数据收集成本较高，跨场景预训练可以利用已有的多场景数据提高模型的初始性能，从而减少在目标场景上所需的训练数据。

Related Work

第一种是Lumigraph and light field representation，directly synthesize novel views by interpolating the input images but require very dense scene capture

第二种是Layered depth images，work for sparse input views but rely on depth maps or estimated depth with sacrificed quality.

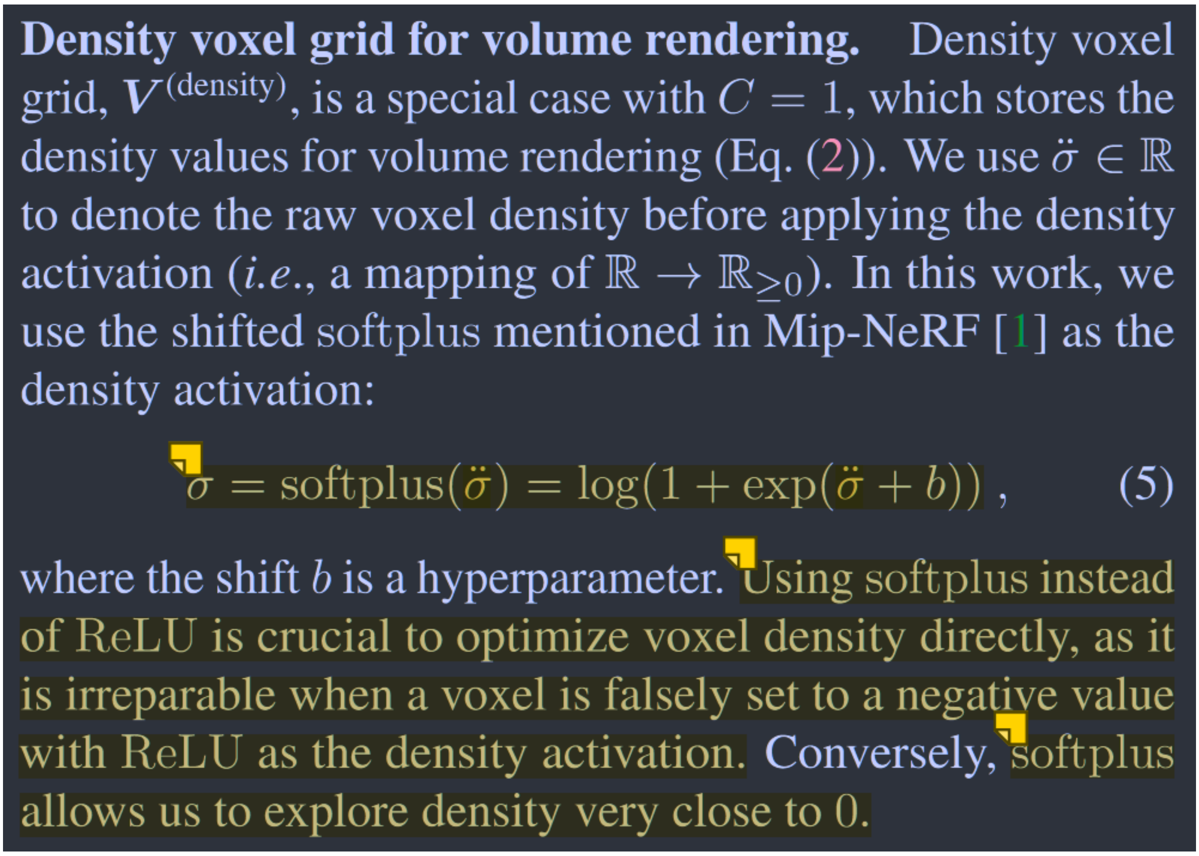
第三种是Recent approaches employ 2D/3D Convolutional Neural Network (CNNs) to estimate multiplane images (MPIs) for forward-facing captures; estimate voxel grid [17, 32, 49] for inward-facing captures.

第四种是比较像的hybrid

NSVF，Unfortunately, gradient-based optimization is not directly applicable to their methods due to their topological data structures or the lack of priors

Post-activated density voxel grid

> A voxel-grid representation models the modalities of interest (e.g., density, color, or feature) explicitly in its grid cells.



Density voxel grid for volume rendering. Density voxel grid, $V^{(\text{density})}$, is a special case with $C = 1$, which stores the density values for volume rendering (Eq. (2)). We use $\ddot{\sigma} \in \mathbb{R}$ to denote the raw voxel density before applying the density activation (*i.e.*, a mapping of $\mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$). In this work, we use the shifted softplus mentioned in Mip-NeRF [1] as the density activation:

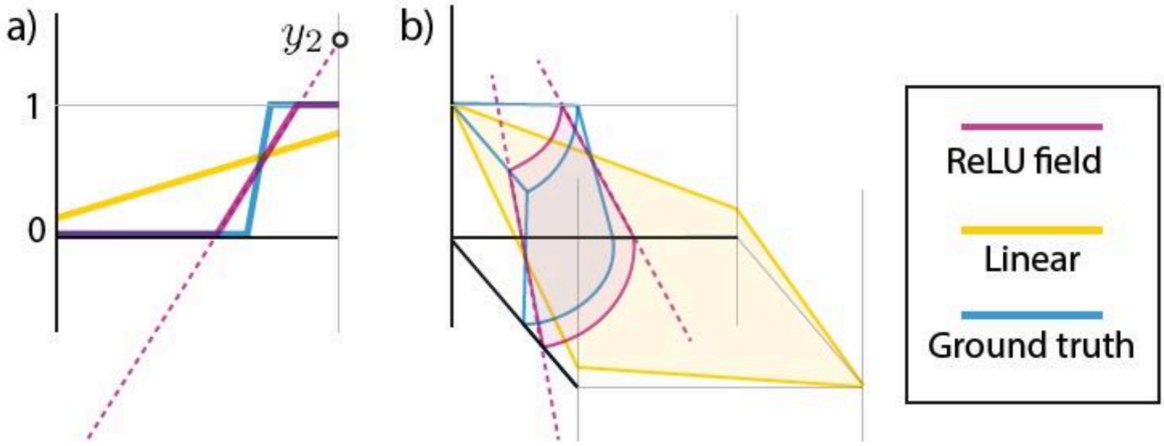
$$\sigma = \text{softplus}(\ddot{\sigma}) = \log(1 + \exp(\ddot{\sigma} + b)) \,, \qquad (5)$$

where the shift b is a hyperparameter. Using softplus instead of ReLU is crucial to optimize voxel density directly, as it is irreparable when a voxel is falsely set to a negative value with ReLU as the density activation. Conversely, softplus allows us to explore density very close to 0.

作者认为softplus比ReLU好的点在于

- 不会像ReLU那样把密度值为负时，强硬的设置为0，softplus更有连续性
- softplus也可以把值设置得非常接近0

我个人理解是引入 SoftPlus/ReLU 之后，网格顶点的值可以学的很广（相邻顶点可以出现很大的区别，如下图左图中的 y_1, y_2 ），增强了网格拟合高频信号的能力。感觉有点像位置编码的作用？反正都是将高频信息变得更加容易学习。



那么问题来了，这个softplus和三线插值以及alpha应该放在什么位置呢？

作者提出了三种方式来实验

$$\alpha^{(\text{pre})} = \underline{\text{interp}}\left(\boldsymbol{x}, \underline{\text{alpha}}\left(\underline{\text{softplus}}\left(\boldsymbol{V}^{(\text{density})}\right)\right)\right) \quad , \tag{6a}$$

$$\alpha^{(\text{in})} = \underline{\text{alpha}}\left(\underline{\text{interp}}\left(\boldsymbol{x}, \underline{\text{softplus}}\left(\boldsymbol{V}^{(\text{density})}\right)\right)\right) \quad , \tag{6b}$$

$$\alpha^{(\text{post})} = \underline{\text{alpha}}\left(\underline{\text{softplus}}\left(\underline{\text{interp}}\left(\boldsymbol{x}, \boldsymbol{V}^{(\text{density})}\right)\right)\right) \quad . \tag{6c}$$

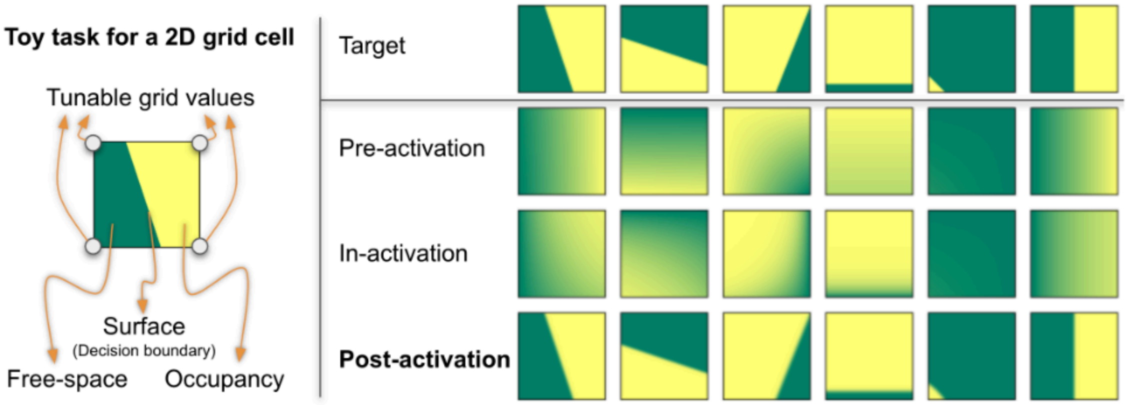
其中alpha是这个

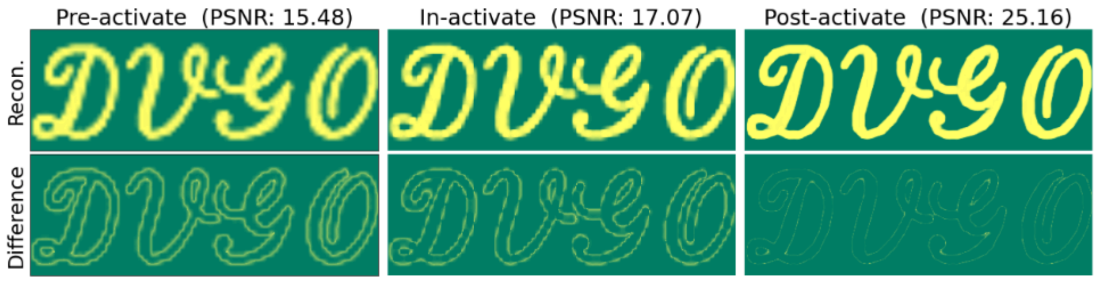
$$\hat{C}(\boldsymbol{r}) = \left(\sum_{i=1}^{K} T_i \alpha_i \boldsymbol{c}_i\right) + T_{K+1}\boldsymbol{c}_{\text{bg}} \ , \tag{2a}$$

$$\alpha_i = \text{alpha}(\sigma_i, \delta_i) = 1 - \exp(-\sigma_i \delta_i) \ , \tag{2b}$$

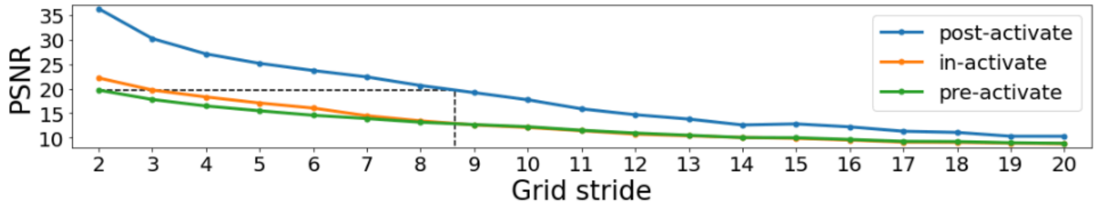
$$T_i = \prod_{j=1}^{i-1}(1 - \alpha_j) \ , \tag{2c}$$

然后作者发现post的方式最好，称为produce sharp surface。作者认为把先插值、后非线性可能有什么神奇的功效。这是他的对比图




Pre-activate (PSNR: 15.48)
In-activate (PSNR: 17.07)
Post-activate (PSNR: 25.16)
Recon.
Difference

(a) Visual comparison of image fitting results under grid resolution $(H/5) \times (W/5)$. The first row is the results of pre-, in-, and post-activation. The second row is their per-pixel absolute difference to the target image.

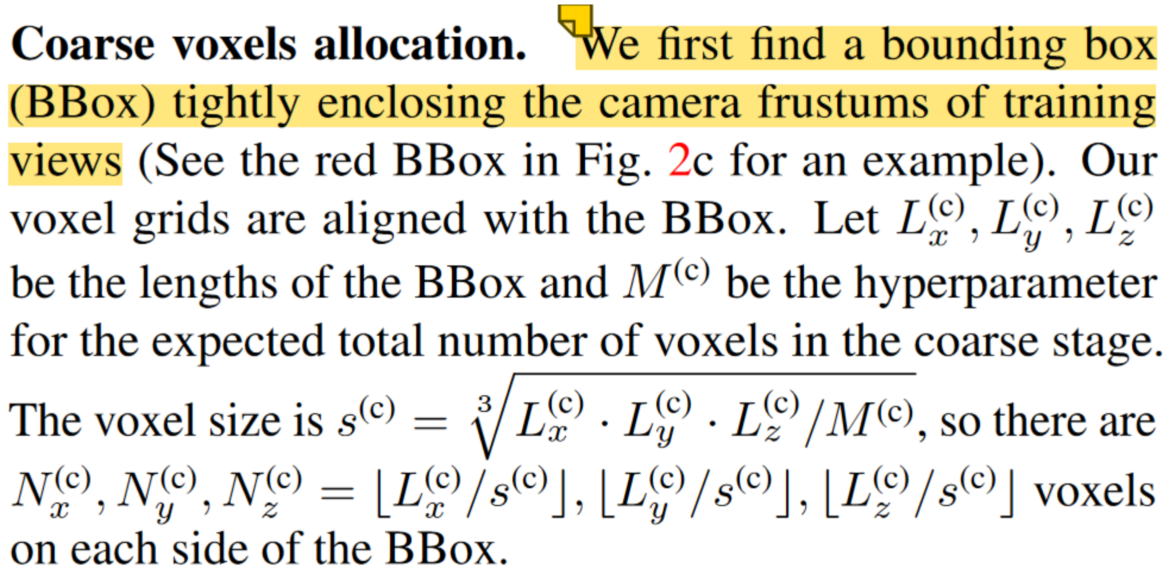


(b) PSNRs achieved by pre-, in- and post-activation under different grid strides. A grid stride s means that the grid resolution is $(H/s) \times (W/s)$. The black dashed line highlights that post-activation with stride ≈ 8.5 can achieve the same PSNR as pre-activation with stride 2 in this example.

所以就选了post的方式。

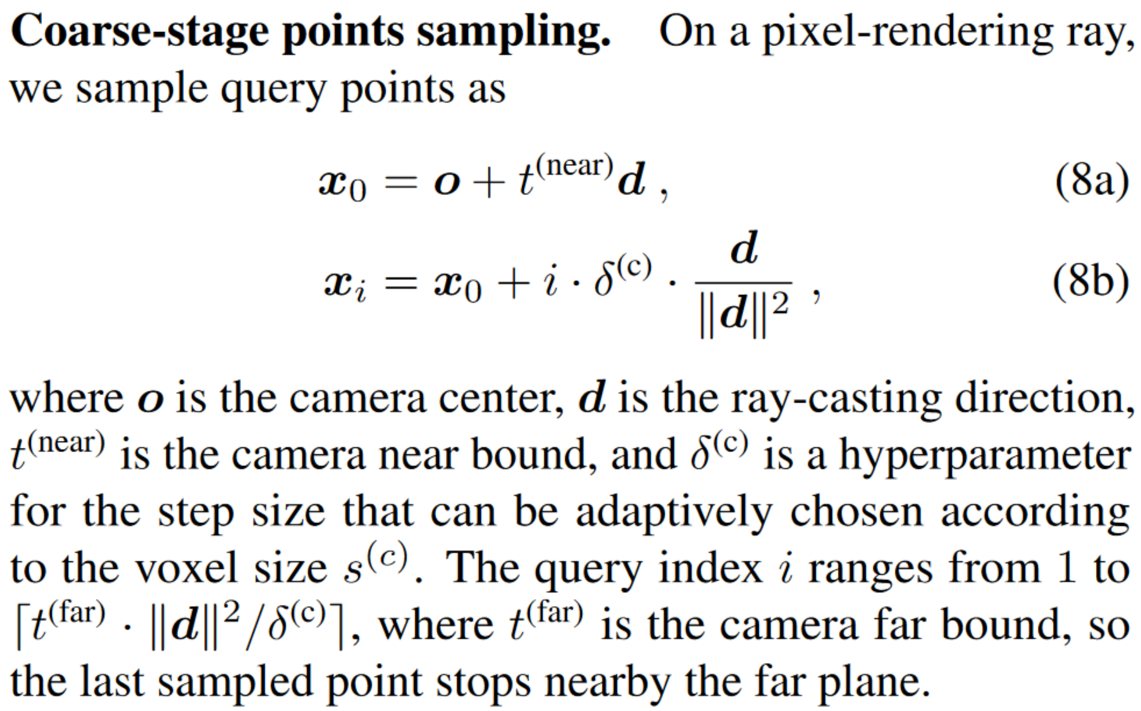
Coarse geometry searching

Coarse voxels allocation



Coarse voxels allocation. We first find a bounding box (BBox) tightly enclosing the camera frustums of training views (See the red BBox in Fig. 2c for an example). Our voxel grids are aligned with the BBox. Let $L_x^{(\mathrm{c})}, L_y^{(\mathrm{c})}, L_z^{(\mathrm{c})}$ be the lengths of the BBox and $M^{(\mathrm{c})}$ be the hyperparameter for the expected total number of voxels in the coarse stage. The voxel size is $s^{(\mathrm{c})} = \sqrt[3]{L_x^{(\mathrm{c})} \cdot L_y^{(\mathrm{c})} \cdot L_z^{(\mathrm{c})} / M^{(\mathrm{c})}}$, so there are $N_x^{(\mathrm{c})}, N_y^{(\mathrm{c})}, N_z^{(\mathrm{c})} = \lfloor L_x^{(\mathrm{c})}/s^{(\mathrm{c})} \rfloor, \lfloor L_y^{(\mathrm{c})}/s^{(\mathrm{c})} \rfloor, \lfloor L_z^{(\mathrm{c})}/s^{(\mathrm{c})} \rfloor$ voxels on each side of the BBox.

Coarse-stage points sampling



Coarse-stage points sampling. On a pixel-rendering ray, we sample query points as

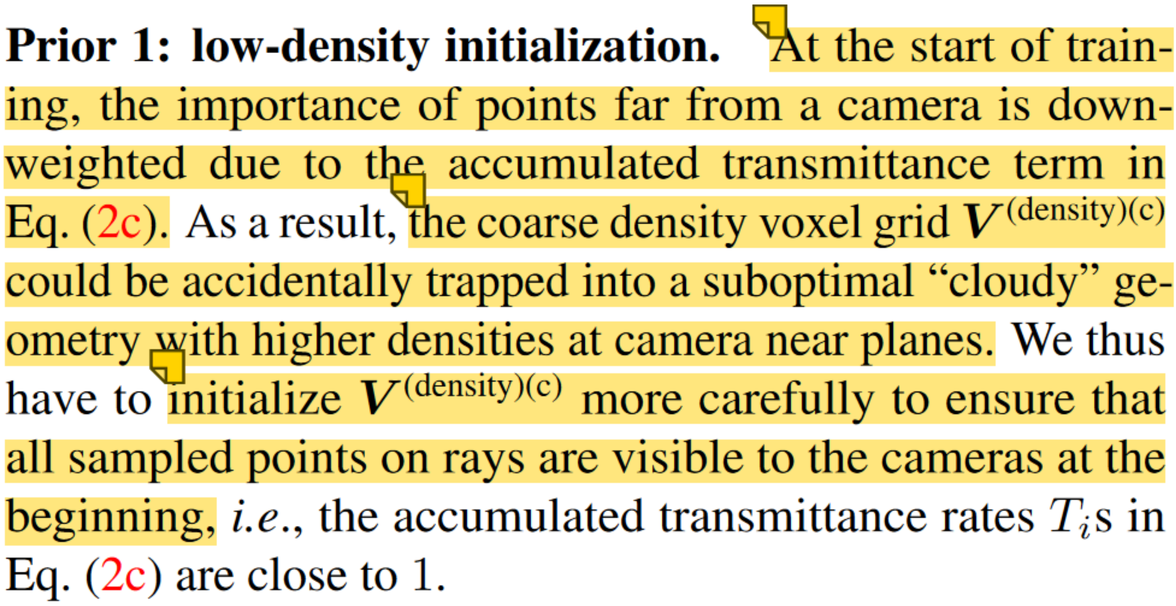
$$\boldsymbol{x}_0 = \boldsymbol{o} + t^{(\mathrm{near})}\boldsymbol{d} \,, \tag{8a}$$

$$\boldsymbol{x}_i = \boldsymbol{x}_0 + i \cdot \delta^{(\mathrm{c})} \cdot \frac{\boldsymbol{d}}{\|\boldsymbol{d}\|^2} \,, \tag{8b}$$

where $\boldsymbol{o}$ is the camera center, $\boldsymbol{d}$ is the ray-casting direction, $t^{(\mathrm{near})}$ is the camera near bound, and $\delta^{(\mathrm{c})}$ is a hyperparameter for the step size that can be adaptively chosen according to the voxel size $s^{(c)}$. The query index i ranges from 1 to $\lceil t^{(\mathrm{far})} \cdot \|\boldsymbol{d}\|^2 / \delta^{(\mathrm{c})} \rceil$, where $t^{(\mathrm{far})}$ is the camera far bound, so the last sampled point stops nearby the far plane.

在 coarse 阶段，作者仅对 coarse grid 进行优化，其中有 coarse density grid $\mathbf{V}^{(\mathrm{density})(\mathrm{c})}$ 和 coarse color grid $\mathbf{V}^{(\mathrm{rgb})(\mathrm{c})}$ 。这一阶段主要是 coarse density grid 的学习。注意，这个 coarse color grid 是 view-invariant，可以理解成漫反射项，这对 coarse 来说足够了，可以减轻训练的难度。

作者发现如果不对存取 density 的 grid 加以约束，则会很容易陷入局部最优解（集中在相机近平面），具体阐述如下。



Prior 1: low-density initialization. At the start of training, the importance of points far from a camera is down-weighted due to the accumulated transmittance term in Eq. (2c). As a result, the coarse density voxel grid $\boldsymbol{V}^{(\text{density})(\text{c})}$ could be accidentally trapped into a suboptimal "cloudy" geometry with higher densities at camera near planes. We thus have to initialize $\boldsymbol{V}^{(\text{density})(\text{c})}$ more carefully to ensure that all sampled points on rays are visible to the cameras at the beginning, *i.e.*, the accumulated transmittance rates T_is in Eq. (2c) are close to 1.

因此作者提出了两个 prior 来避免这一情况。

一是初始化的时候给较低的 density。由于 σ 的计算如下：

$$\sigma = \text{softplus}(\ddot{\sigma}) = \log(1 + \exp(\ddot{\sigma} + b))$$

其中 $\ddot{\sigma}$ 初始化为 0， b 则为：

$$b = \log\left((1 - \alpha^{(\text{init})(\text{c})})^{-\frac{1}{s^{(\text{c})}}} - 1\right)$$

其中 $\alpha^{(\text{init})(\text{c})}$ 为超参数。

二是提出了基于视角计数的学习率调整策略 (view-count-based learning rate)。统计每个 grid 被训练样本视角看到的次数 n_j 然后设置该 grid的学习率为 $n_j/n_{\max}$ ，其中 $n_{\max}$ 为可视次数最多的 grid 的可视次数。

Fine Detail Reconstruction

在 Coarse 阶段结束后，DVGO 就得到了一个大概能够正确描述目标物体的 corase density grid $\mathbf{V}^{(\text{density})(\text{c})}$ 了。

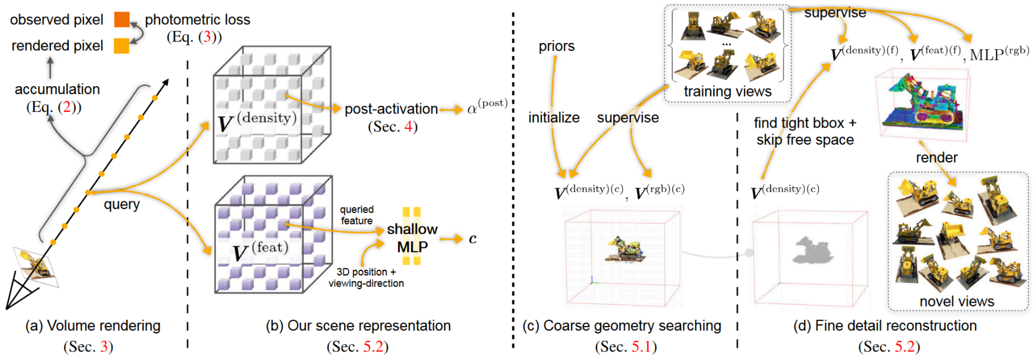
有了 $\mathbf{V}^{(\text{density})(\text{c})}$ 之后，作者一方面用来加大采样密度，另一方面做跳点加速（剪枝无效空间）。

之后作者会进行一个类似 Plenoxel 的上采样的过程，即用 coarse density grid $\mathbf{V}^{(\text{density})(\text{c})}$ 三线性插值得到 fine density grid $\mathbf{V}^{(\text{density})(\text{f})}$ ，从而实现更加准确的 density 估计（注意 $\mathbf{V}^{(\text{density})(\text{f})}$ 也会参与跳点）。

在 color 建模方面，相比于 Plenoxel 基于球谐函数建模 color，DVGO 还是选择基于 MLP 的隐式方式来预测，只是利用网格顶点来存点的特征，然后对样本点利用三线性插值得到对应特征后结合 MLP 来预测一个和视角相关的 color。

流程图

总体设计如下



(a) Volume rendering (Sec. 3)

(b) Our scene representation (Sec. 5.2)

(c) Coarse geometry searching (Sec. 5.1)

(d) Fine detail reconstruction (Sec. 5.2)

Figure 2. **Approach overview.** We first review NeRF in Sec. 3. In Sec. 4, we present a novel post-activated density voxel grid to support sharp surface modeling in lower grid resolutions. In Sec. 5, we show our approach to the reconstruction of radiance field with super-fast convergence, where we first find a coarse geometry in Sec. 5.1 and then reconstruct the fine details and view-dependent effects in Sec. 5.2.