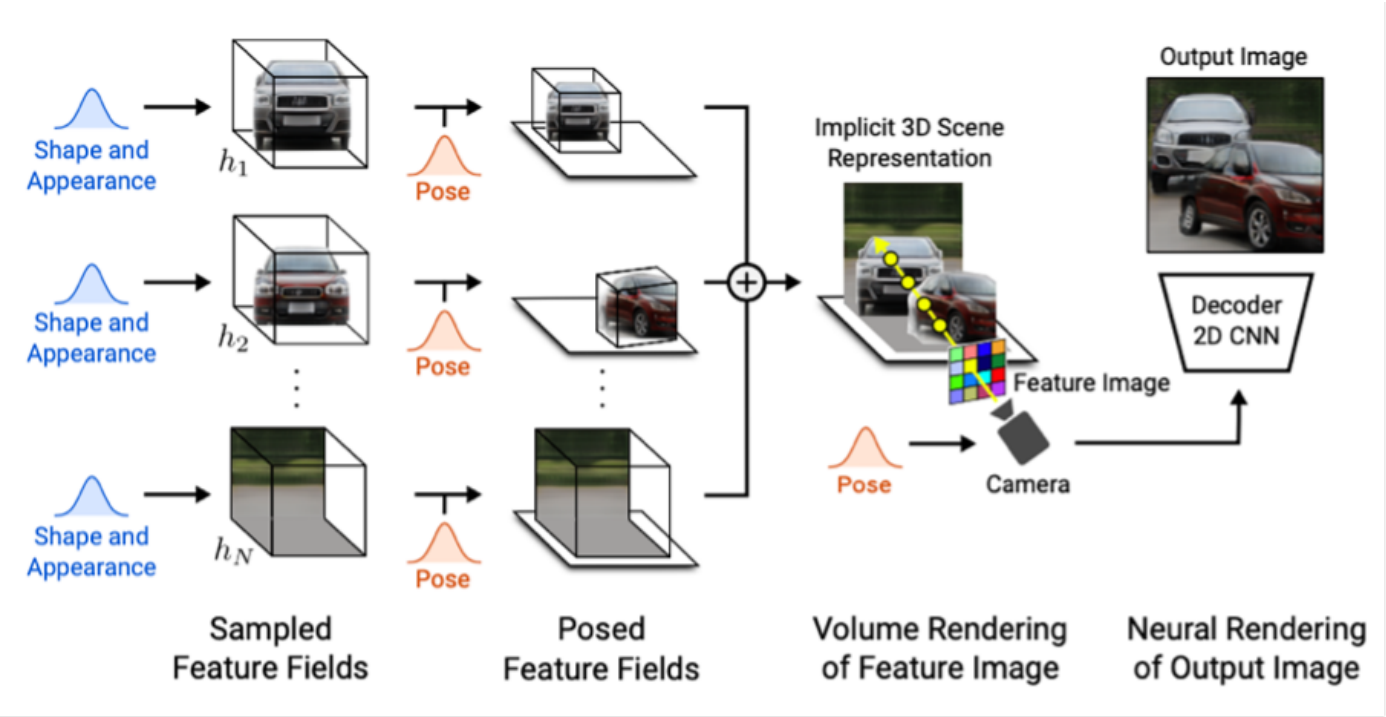
新视角图像生成：讨论基于NeRF的泛化方法



新视角图像生成（NVS）是计算机视觉的一个应用领域，在1998年SuperBowl的比赛，CMU的RI曾展示过给定多摄像头立体视觉（MVS）的NVS，当时这个技术曾转让给美国一家体育电视台，但最终没有商业化；英国BBC广播公司为此做过研发投入，但是没有真正产品化。

在基于图像渲染（IBR）领域，NVS应用有一个分支，即基于深度图像的渲染（DBIR）。另外，在2010年曾很火的3D TV，也是需要从单目视频中得到双目立体，但是由于技术的不成熟，最终没有流行起来。当时基于机器学习的方法已经开始研究，比如Youtube曾经用图像搜索的方法来合成深度图。

几年前我曾介绍过深度学习在NVS的应用：

null

最近一段时间，神经辐射场（NeRF）已经成为表示场景和合成照片逼真图像的有效范例，其最直接的应用就是NVS。传统NeRF的一个主要限制是，通常无法在训练视点显著不同的新视点生成高质量的渲染。下面以此展开讨论NeRF的泛化方法，这里忽略基础的NeRF原理介绍。有兴趣的请参考综述论文：

null

null

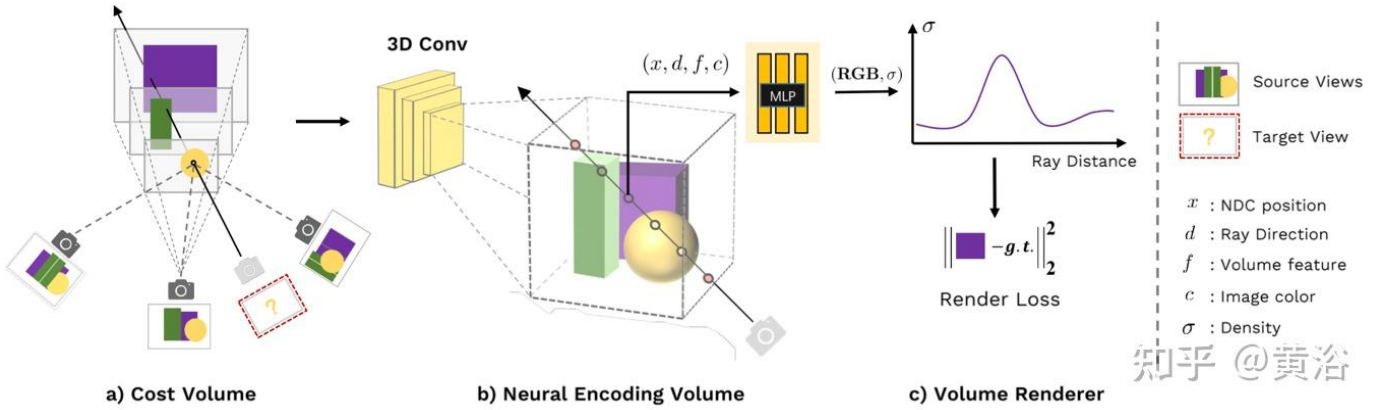
论文【2】提出了一种通用的深度神经网络MVSNeRF，实现跨场景泛化，推断从仅三个附近的输入视图重建辐射场。该方法利用平面扫描成本体（广泛用于多视图立体视觉）进行几何感知场景推理，并与基于物理的体渲染相结合，用于神经辐射场重建。

该方法利用深度MVS的成功，在成本体上应用3D卷积来训练用于3D重建任务的可泛化神经网络。与MVS方法不同的是，MVS方法仅对这样的成本体进行深度推断，而该网络对场景几何和外观进行推理，并输出神经辐射场，从而实现视图合成。具体而言，利用3D CNN，重建（从成本体）神经场景编码体，由编码局部场景几何和外观信息的体素神经特征组成。然后，多层感知器（MLP）在编码体内用三线性插值的神经特征对任意连续位置处的体密度和辐射度进行解码。本质上，编码体是辐射场的局部神经表征；其一旦估计，可直接用于（丢弃

3D CNN）可微分光线行进（ray-marching）进行最终渲染。

与现有的MVS方法相比，MVSNeRF启用可微分神经渲染，在无3D监督的情况下进行训练，并优化推断时间，以进一步提高质量。与现有的神经渲染方法相比，类似MVS的体系结构自然能够进行跨视图的对应推理，有助于对未见测试场景进行泛化，引向更好的神经场景重建和渲染。

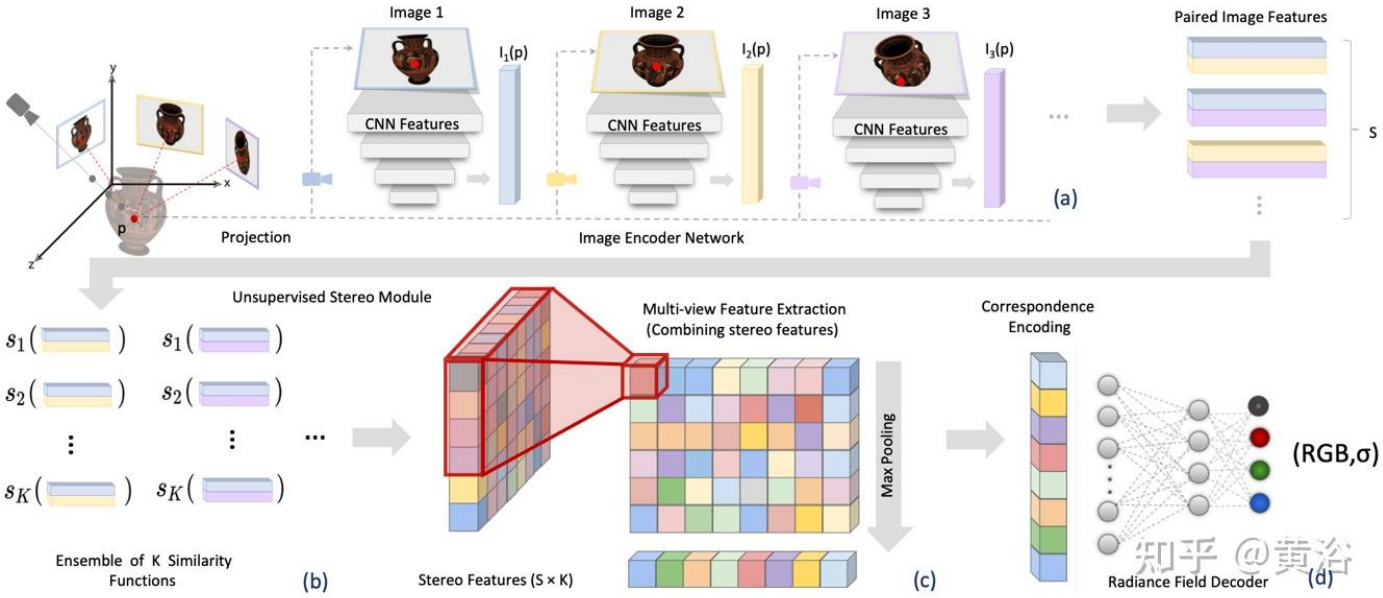
如图1是MVSNeRF的概览：（a）基于摄像头参数，首先将2D图像特征warp（单应变换）到一个平面扫描（plane sweep）上，构建成本体；这种基于方差的成本体编码了不同输入视图之间的图像外观变化，解释了由场景几何和视图相关明暗效果引起的外观变化；（b）然后，用3D CNN重建逐体素神经特征的一个神经编码体；3D CNN 是一个3D UNet，可以有效地推断和传播场景外观信息，从而产生有意义的场景编码体；注：该编码体是无监督预测的，并在端到端训练中用体渲染进行推断；另外，还将原图像像素合并到下一个体回归阶段，这样可恢复下采样丢失的高频；（c）用MLP，通过编码体插值的特征，在任意位置回归体密度和RGB辐射度，这些体属性由可微分光线行进做最终的渲染。



论文【3】提出立体视觉辐射场（SRF），一种端到端训练的神经视图合成方法，可泛化到新场景，并且在测试时只需要稀疏视图。其核心思想是一种受经典多视图立体视觉（MVS）方法启发的神经架构，在立体图像中找到相似的图像区域来估计表面点。输入编码器网络10个视图，提取多尺度特征。多层感知器（MLP）替换经典的图像块或特征匹配，输出相似性分数的集成。在SRF中，每个3D点给定输入图像中立体视觉对应的一个编码，预先预测其颜色和密度。通过成对相似性的集成，该编码被隐式地学习——模拟经典立体视觉。

已知摄像头参数，给定一组N个参考图像， SRF预测3D点的颜色和密度。构造SRF模型f，类似于经典的多视图立体视觉方法：（1）为了编码点的位置，将其投影到每个参考视图中，并构建局部特征描述符；（2） 如果在一个表面上并且照片一致，，特征描述符应该互相匹配；用一个学习的函数模拟特征匹配，对所有参考视图的特征进行编码；（3） 该编码由一个学习的解码器进行解码，成为NeRF表征。如图2给出SRF的概览：（a）提取图像特征；（b）通过一个学习的相似度函数

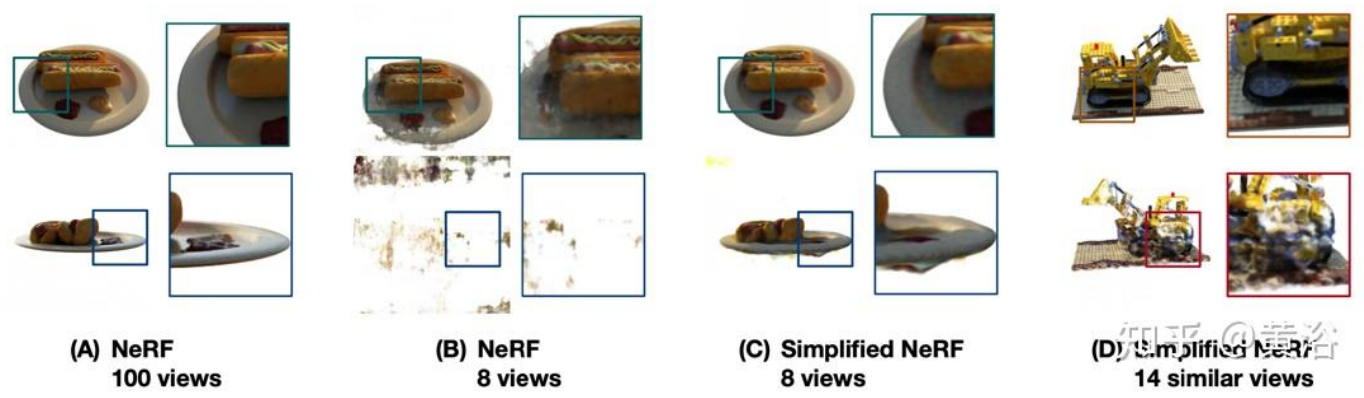
模拟寻找照片一致性的过程，得到一个立体特征矩阵（SFM）；（c）聚集信息，获取多视图特征矩阵（MFM）；（d）最大池化获取对应和颜色的紧凑编码，解码后得到颜色和体密度。



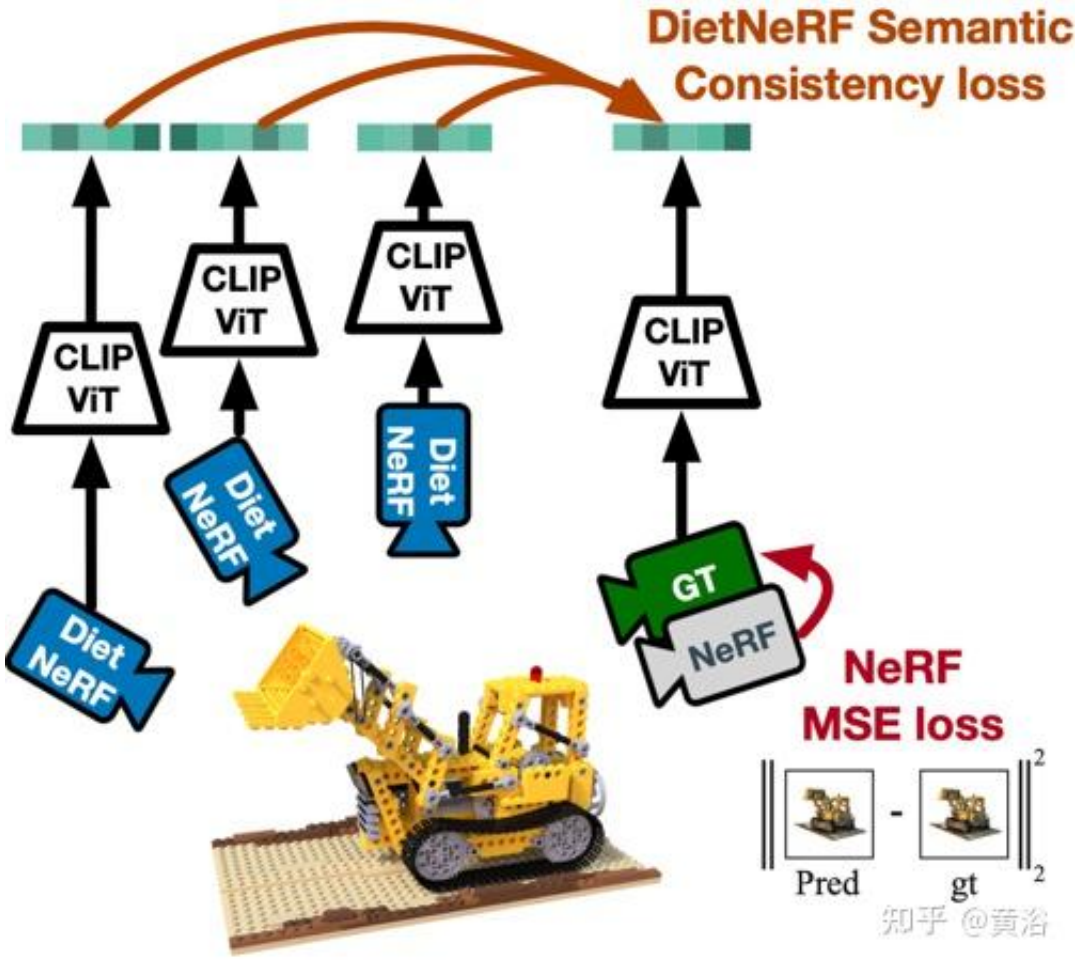
论文【4】提出DietNeRF，一个从几个图像估计的3D神经场景表征。其引入一种辅助语义一致性损失，鼓励新姿态进行真实的渲染。

当NeRF只有少数视图可用时，渲染问题是未约束的；除非严格正则化，否则NeRF通常会出现退化解。如图3所示：（A）从均匀采样的姿态中对一个目标进行了100次观察时，NeRF估计一个详细而准确的表征，允许纯粹从多视图一致性进行高质量视图合成；（B） 在只有8个视图的情况下，将目标放置在训练摄像头的近场中，相同的NeRF过拟合，导致在训练摄像头附近的姿态出现目标错位，并退化；（C） 当正则化、简化、调整和手工重新初始化时，NeRF可以收敛，但不再捕获

精细细节；（D） 如果没有关于类似目标的先验知识，单场景视图合成无法合理地完成未观察区域。



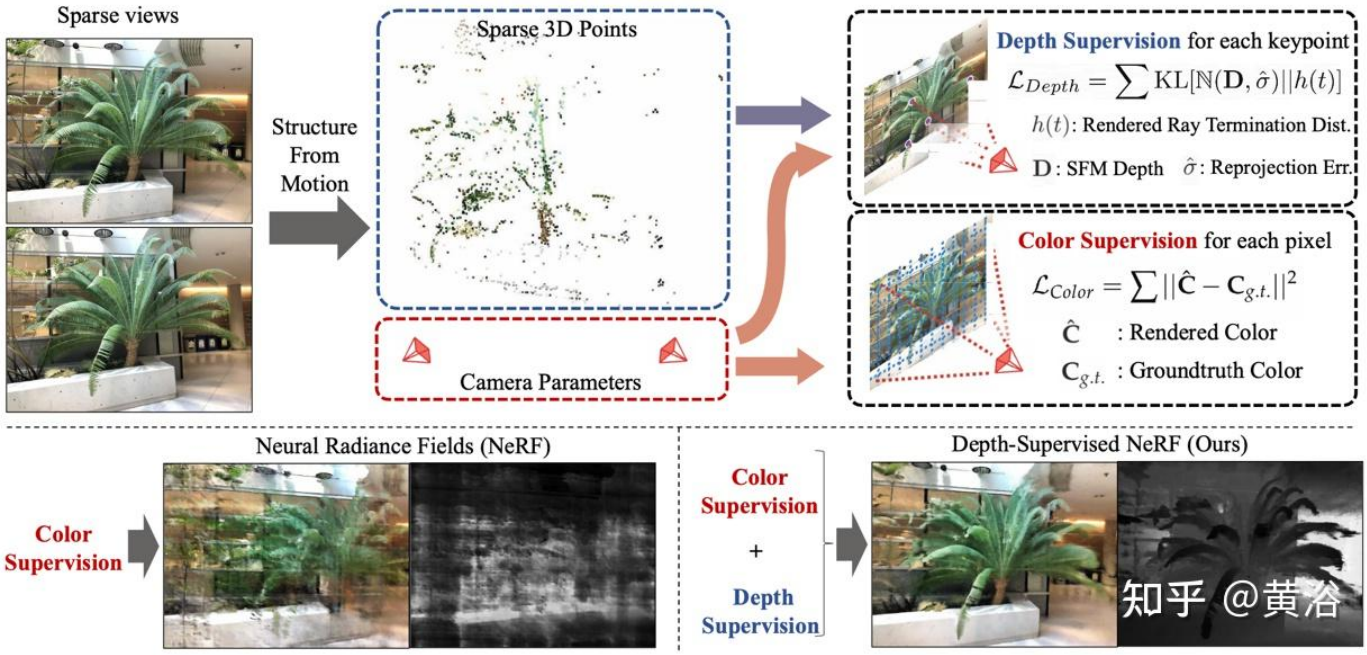
如图4是DietNeRF工作示意图：基于“从任何角度看，一个物体都是那个物体”的原则，DietNeRF监控任意姿态的辐射场（DietNeRF摄像头）；计算语义一致性损失，是在捕获高级场景属性的特征空间中，而不是在像素空间中；所以用CLIP这个视觉Transformer提取渲染的语义表征，然后最大化与真值视图表征的相似性。



实际上，单视图2D图像编码器学习的场景语义先验知识，就可以约束一个3D表征。DietNeRF在自然语言监督下，从网络挖掘的数亿单视图2D照片集进行训练：（1）给定来自相同姿态的给定输入视图，可正确地渲染，（2）不同随机姿态下匹配高级语义属性。语义损失函数能够从任意姿态监督DietNeRF模型。

论文【5】提出DS-NeRF，采用一种学习辐射场的损失，利用现成的深度图监督，如图5所示。有这样一个事实，即当前的NeRF流水线需要具有已知摄像头姿态的图像，这些姿态通常通过运动恢复结构（SFM）来估计。至关重要的是，SFM还产生了稀疏的3D点，在训练期间用作“自由”深度监督：增加一个

损失，鼓励一个光线的终止深度分布与一个给定的3D关键点相匹配，包括深度不确定性。

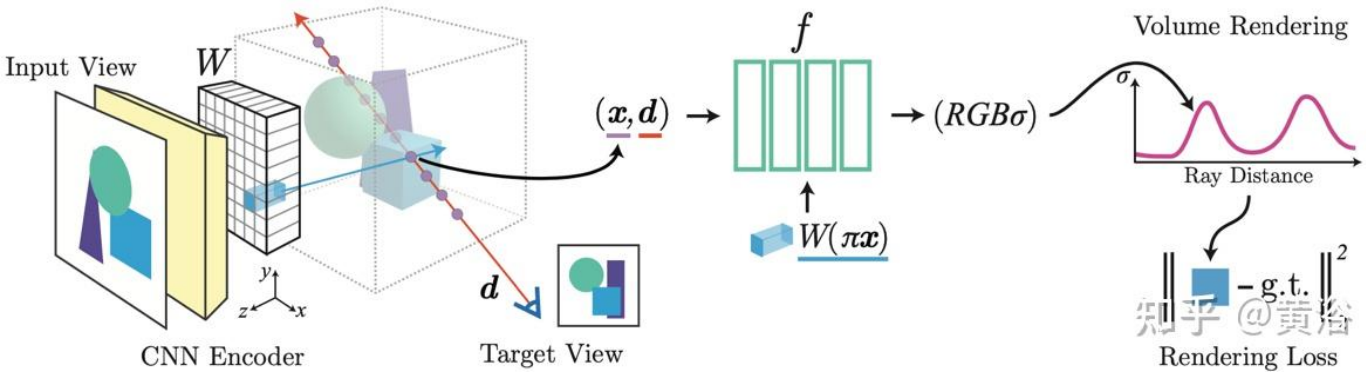


论文【6】提出pixelNeRF，一个基于一或多输入图像预测连续神经场景表征的学习框架。其引入一种全卷积方式在图像输入上调节NeRF架构，使得网络能够跨多场景进行训练来学习一个场景的先验知识，从而能够从稀疏的一组视图（最少就一个）以前馈方式进行新视图合成。利用NeRF的体渲染方法，pixelNeRF可以直接从图像中训练，无需额外的3D监督。

具体地讲，pixelNeRF首先从输入图像计算全卷积图像特征网格（feature grid），在输入图像上调节NeRF。然后，对于视图坐标系中感兴趣的每个3D查询空间点x和视图方向d，通过投影和双线性插值采样相应的图像特征。查询规范与图像特征一起发送到输出密度和颜色的NeRF网络，其中空间图像特征作为一个残差馈送到每个层。当有多个图像可用时，首先将输

入编码为每个摄像头坐标系的潜表征，在预测颜色和密度之前将其合并在中间层中。该模型训练基于一个真值图像和一个体渲染视图之间的重建损失。

pixelNeRF框架如图6所示：对于沿视图方向**d**、一个目标摄像头光线的一个3D查询点x，通过投影和插值从特征体**W** 提取对应的图像特征；然后将该特征与空间坐标一起传递到NeRF网络*f*中；输出RGB和密度值被用于体渲染，并与目标像素值进行比较；坐标x和**d** 在输入视图的摄像头坐标系中。



可以看出，PixelNeRF和SRF用从输入图像提取的局部CNN特征，而MVSNeRF通过image warping获得3D成本体，然后由3D CNN处理。这些方法需要许多不同场景的多视图图像数据集进行预训练，获取成本可能很高。此外，尽管预训练阶段很长，但大多数方法都需要在测试时微调网络权重，并且当测试域发生变化时，新视图的质量很容易下降。

当然，DS-NeRF增加额外的深度监督来提高重建精度。Diet-NeRF比较了CLIP在低分辨率下渲染的未见视点嵌入。这种语义一致性损失只能提供高级信息，不能改善稀疏输入的场景几何。

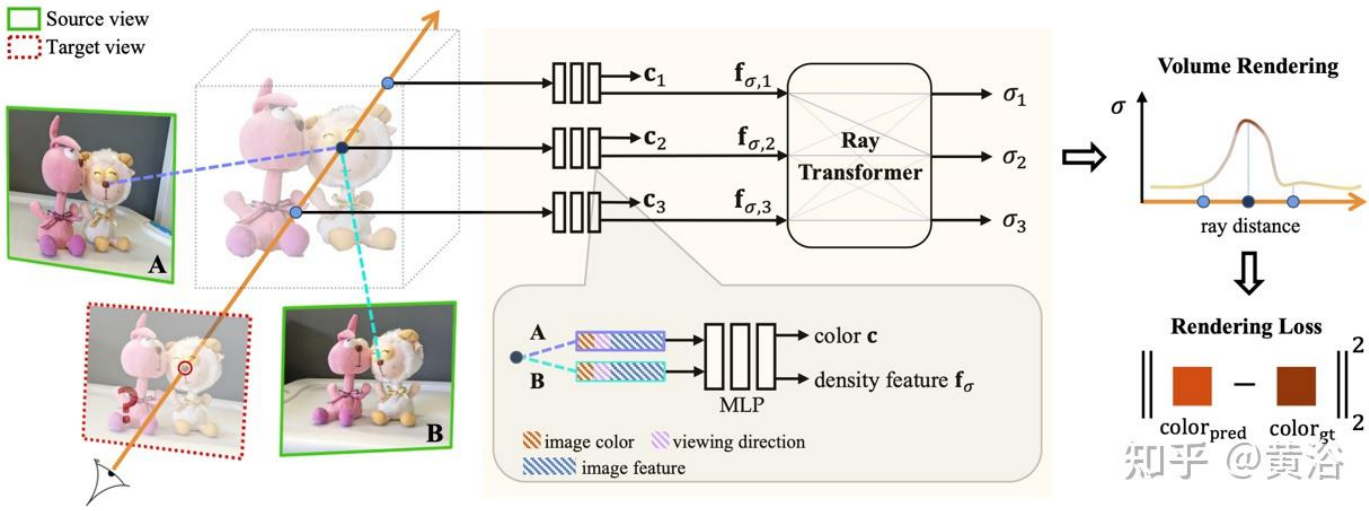
论文【7】提出的IBRNet，其核心包括MLP和光线transformer（经典Transformer架构：位置编码和自注意），用于估计连续5D位置（3D空间位置和2D观看方向）的辐射度和体密度，并从多个源视图实时渲染外观信息。

在渲染时，该方法可以追溯到经典的**基于图像渲染**（**IBR**）工作。不同于神经场景表征，其为渲染优化每个场景函数，IBRNet学习一种通用的视图插值函数，可泛化到新场景。还是经典的体渲染来合成图像，其完全可微分，并且用多视图姿态图像作为监督来训练。

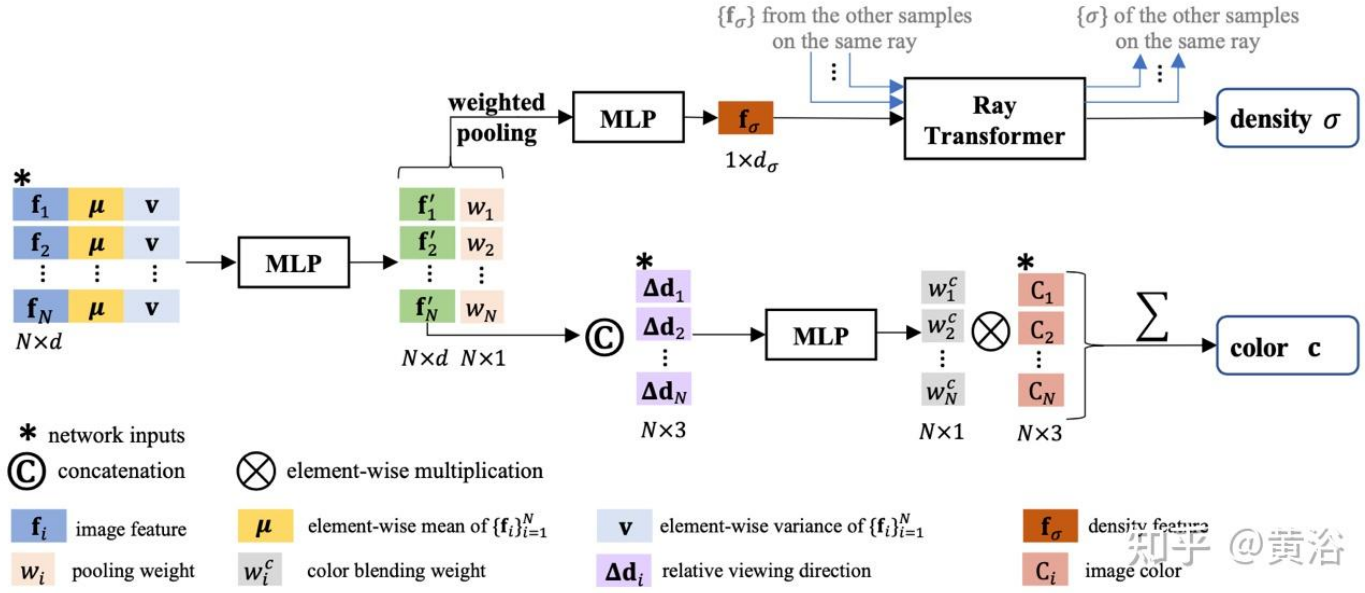
光线transformer沿整个光线考虑这些密度特征来计算每个样本的标量密度值，实现更大空间尺度上的可见性推理（visibility reasoning）。单独地，一个颜色调和（color blending）模块用2D特征和源视图的视线向量导出每个样本的视图相关颜色。最后，体渲染为每条光线计算最终颜色值。

如图7是IBRNet概览：1） 为渲染目标视图（标记“？”图像），首先识别一组相邻的源视图（例如，标记为A和B的视图）并提取图像特征；2） 然后，对目标视图中的每条光线，用IBRNet（黄色阴影区域）计算沿光线的一组样本颜色和密

度；具体而言，对每个样本从相邻源视图中聚合相应的信息（图像颜色、特征和观看方向），生成其颜色**c** 和密度特征；然后，将ray transformer应用于光线上所有样本的密度特征，预测密度值。3） 最后，用体渲染沿光线累积颜色和密度。在重建图像颜色上，可进行端到端的L2损失训练。



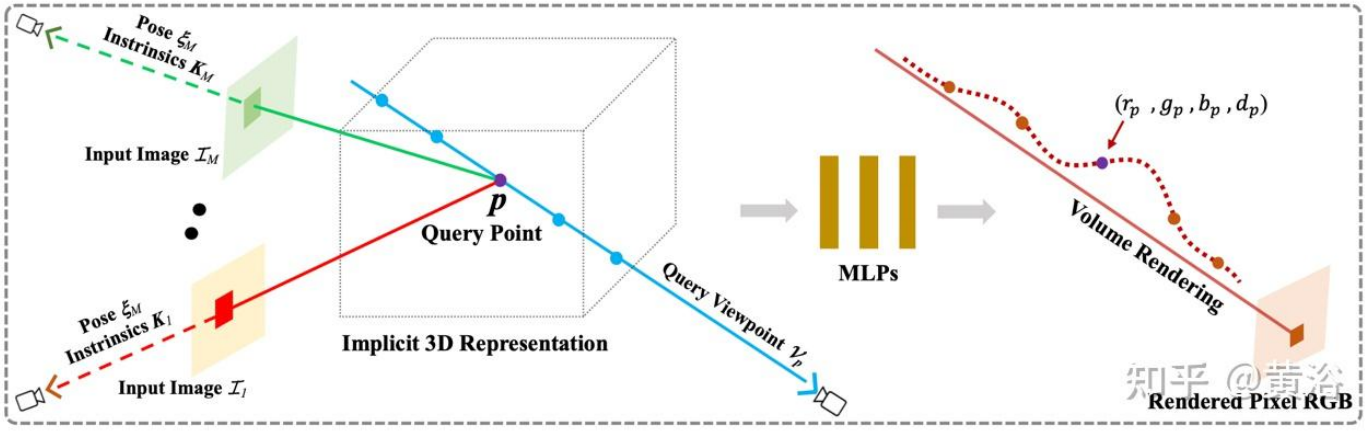
如图8是IBRNet用于连续5D位置的颜色+体密度预测工作：首先将所有源视图中提取的2D图像特征输入到类似PointNet的MLP中，聚合局部和全局信息，产生多视图感知特征和池化权重，用权重来集中特征，进行多视图可见性推理，获得密度特征；这里没有直接从预测单个5D样本的密度σ，而是用ray transformer模块聚集沿光线的所有样本信息；ray transformer模块为光线上的所有样本获取密度特征，并预测其密度；ray transformer模块能够在更长的范围进行几何推理，并改进密度预测；对于颜色预测，将多视图感知特征，与查询光线相对于源视图的观看方向，连接输入一个小网络预测一组调和权重，输出颜色**c** 是源视图的图像颜色加权平均。



这里补充一点：与采用绝对观看方向的NeRF不同，IBRNet考虑相对于源视图的观看方向，即**d** 和 d_i之间的差异，$\Delta d = d - d_i$。 Δd较小，通常意味着目标视图的颜色与源视图*i*相应颜色相似的可能性较大，反之亦然。

论文【8】提出的通用辐射场（GRF），仅从2D观察中表征和渲染3D目标和场景。该网络将3D几何建模为一个通用辐射场，以一组2D图像、摄像机外参姿态和内参为输入，为3D空间每个点构建内部表征，然后渲染从任意位置观察的相应外观和几何。其关键是学习2D图像每个像素的局部特征，然后将这些特征投影到3D点，从而生成通用和丰富的点表征。此外，集成一个注意机制来聚合多个2D视图的像素特征，从而隐式地考虑视觉遮挡问题。

如图9是GRF的示意图：GRF将每个3D点*p*投影到*M*个输入图像的每一个，从每个视图收集每个像素的特征，聚集并馈送到MLP，推断出*p*的颜色和体密度。



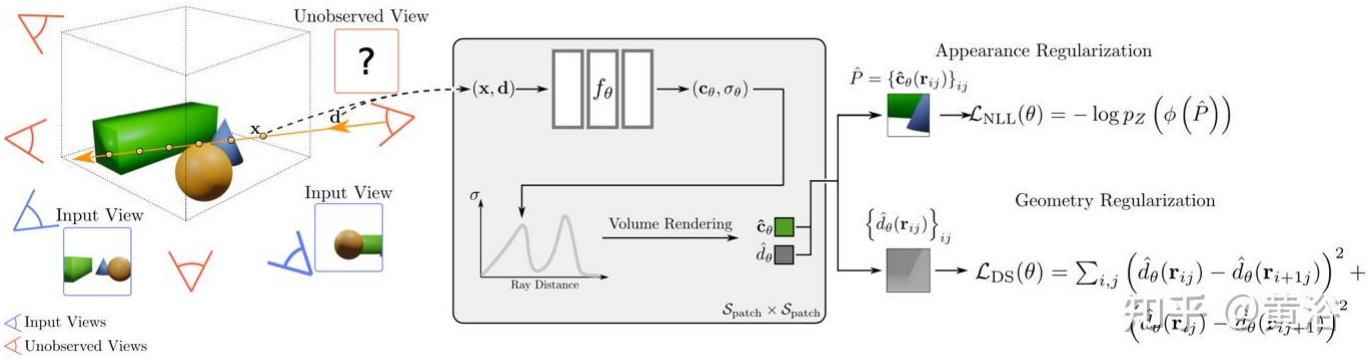
GRF由四部分组成：1）每个2D像素的特征提取器，一个基于CNN的编码器-解码器；2） 2D特征转换为3D空间的重投影；3） 获取3D点通用特征的基于注意聚合器；4）神经渲染器NeRF。

由于没有与RGB图像配对的深度值，因此无法确定像素特征属于哪个特定的3D表面点。在重投影模块中，将像素特征视为3D空间中光线沿线每个位置的表征。形式上，给定一个3D点、一个观察2D视图以及摄像机姿态和内参，相应的2D像素特征可以通过重投影操作进行检索。

在特征聚合器中，注意机制学习所有输入特征的唯一权重，然后聚合在一起。通过一个MLP，3D点的颜色和体密度可以被推断。

论文【9】提出RegNeRF，对未观测视点渲染的图像块几何和外观进行正则化，并在训练期间对光线采样空间进行退火。此外，用归一化流模型正则化未观测视点的颜色。

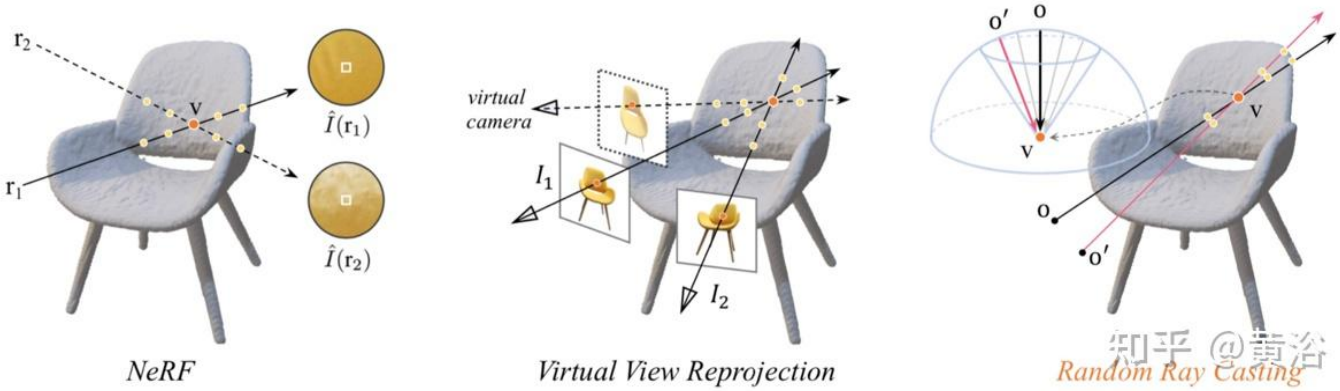
如图10是RegNeRF模型的概览：给定一组输入图像集合（蓝色摄像头），NeRF优化重建损失；然而，对于稀疏输入，这会导致退化解；这项工作对未观察的视图（红色摄像头）进行采样，并正则化从这些视图渲染的图像块几何和外观；更具体地说，对于给定的辐射场，通过场景投射光线，并从未观察的视点渲染图像块；然后，通过训练的归一化流模型，将预测的RGB图像块馈送，并最大化预测的对数似然，从而正则化外观；对渲染的深度图块强制一个平滑度损失，可正则化几何；该方法导致3D一致性表征，甚至对渲染真实新视图的稀疏输入来说，也是如此。



论文【10】研究了一种新视图外推而不是少样本图像合成的方法，即（1）训练图像可以很好地描述目标，（2）训练视点和测试视点的分布之间存在显著差异，其称为RapNeRF（RAy Priors NeRF）。

论文【10】的见解是，3D曲面任意可见投影的固有外观应该是一致的。因此，其提出一种随机光线投射（random ray casting）策略，允许用已见的视图训练未见的视图。此外，根据沿着观测光线的视线方向预先计算的光线图集，可以进一步提高外推视图的渲染质量。一个主要的限制是RapNeRF利用多视图一致性去消除视图强相关效应。

随机光线投射（random ray casting）策略直观解释如图11所示：左图中，有两个观察3-D点v的光线，r1位于训练空间，r2远离训练光线；考虑到NeRF的分布漂移和映射函数 $F_c : (r, f) \rightarrow c$，其沿r2的一些样本辐射将是不精确的；与像素颜色相比，沿r2的辐射累积操作更有可能提供v的反颜色估计；中图是一个简单的虚拟视图重投影，其遵循NeRF公式计算所涉及的像素光线，从训练光线池中找到击中同一3D点的虚拟光线所对应的光线，实践中很不方便；右图中，对于特定的训练光线（从o投射并穿过v），**随机光线投射（RRC）**策略在一个圆锥内随机生成一条未见过的虚拟光线（从o'投射并穿过v），然后基于训练光线在线指定一个伪标签；RRC支持用见过的光线训练未见过的光线。



RRC策略允许以在线方式为随机生成的虚拟光线分配伪标签。具体地说，对于一个训练图像*I*中的一个感兴趣像素，给出其世界坐标系中的观察方向*d*、相机原点*o*和深度值 t_z，并且光线 $r = o + td$。这里，使用预训练的NeRF对t_z预计算和存储。

设 $v = o + t_z d$表示*r*命中的最近3D曲面点。在训练阶段，将*v*视为新原点，并在圆锥内从*v*随机投射一条光线，其中心线为矢量 $\bar{vo} = -t_z d$。这可以轻松实现，只要将 $\bar{vo}$转换到球形空间并引入一些随机干扰Δφ和Δθ到φ和θ。这里，φ和θ分别是$\bar{vo}$的方位角和仰角。Δφ和Δθ从预定义间隔 [−η, η] 均匀采样。由此得到 $\theta' = \theta + \Delta\theta$和$\varphi' = \varphi + \Delta\varphi$。因此，可以从一个随机原点o'投射一个也通过v的虚拟光线。这样，可以将颜色强度 $I(r)$真值视为 $\tilde{I}(r')$的伪标记。

基础NeRF利用“方向嵌入”来编码场景的照明效果。场景拟合过程使得训练的颜色预测MLP严重依赖于视线方向。对于新视图内插，这不是问题。然而，由于训练和测试光线分布之间存在一些差异，这可能不适合于新视图外推。一个天真的想法是直接移除方向嵌入（表示为“NeRF w/o dir”）。然而，这通常会产生伪影图像，如意外的波纹和非平滑的颜色。这意味着光线的观察方向也可能与表面平滑度有关。

论文【10】计算了一个光线图集（ray atlas），并表明它可以进一步提高外插视图的渲染质量，同时不涉及内插视图的问题。光线图集类似于一个纹理图集，但它存储每个3D顶点的全局光线方向。

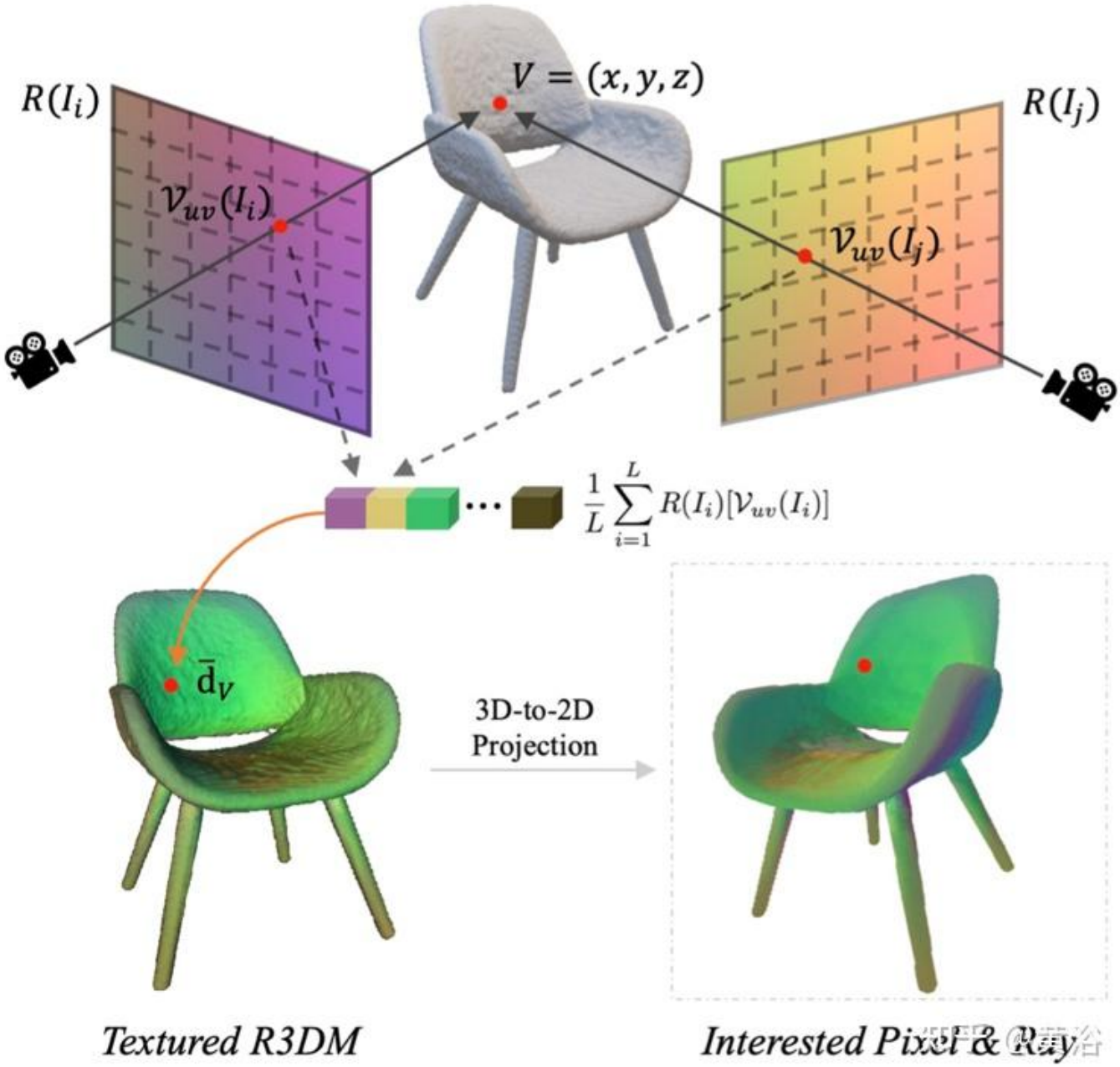
特别是，对于每个图像（例如，图像*I*），对所有空间位置抓取其光线的观察方向，从而生成一个光线图。从预训练的NeRF中提取一个粗糙的3D网格（R3DM），并将光线方向映射到3D顶点。以顶点V=(x,y,z)为例，其全局光线方向 $\bar{dV}$ 应表示为

$$\bar{d}V = \frac{1}{L}\sum_{i=1}^{L} R(I_i)[\mathcal{V}_{uv}(I_i)]$$

$$\mathcal{V}_{uv}(I_i) = \frac{1}{z} K\mathcal{T}_{w2c}(I_i)V$$

其中K是摄像头内参，$\Gamma_{w2c}(I_i)$是图像 I_i的摄像头-世界坐标系转换矩阵，$V_{uv}(I_i)$是顶点V在图像I_i的2-D投影位置，L是在顶点V重建中训练图像数。对于一个任意摄像头姿态的每个像素，投影具有光线图纹理的3D网格（R3DM）到2D可获得一个全局光线先验 $\bar{d}$。

如图12就是光线图集的示意图：即从训练光线中捕获一个光线图集并用之对椅子的粗糙3D网格（R3DM）附加纹理；$R(I_i)$是训练图像I_i的光线图。



在训练RapNeRF时，用感兴趣像素I(r)的$\bar{d}$来替换其在F_c中的d，进行颜色预测。这种替代机制发生的概率为0.5。在测试阶段，样本x的辐射度c近似为：

$$c = F_c(\bar{d}, F_\sigma(x))$$

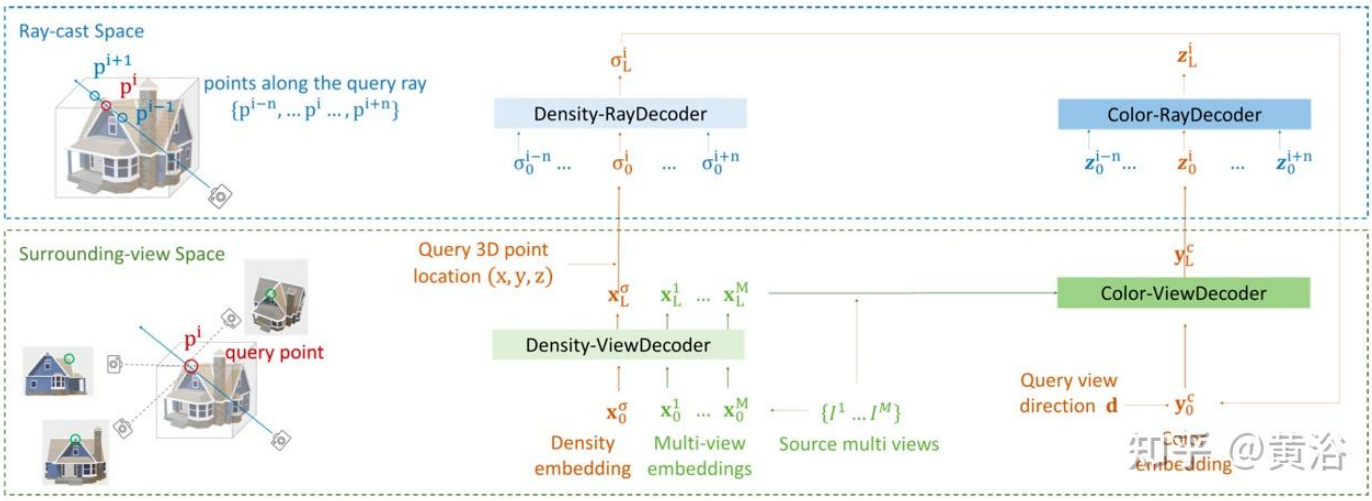
其中映射函数$F_\sigma(x) : x \rightarrow (\sigma, f)$。

原始NeRF独立地优化每个场景表征，不需要探索场景之间的共享信息，而且耗时。为了解决这一问题，研究人员提出了PixelNeRF和MVSNeRF 等模型，这些模型接收多个观察者视图作为条件输入，学习通用的神经辐射场。遵循分而治之（divide-and-conquer）的设计原则，其包括两个独立的组件：用于单个图像的CNN特征提取器和作为NeRF网络的MLP。对于单视图立体视觉，在这些模型中，CNN将图像映射到特征网格，MLP将查询5D坐标及其对应的CNN特征映射到单个体密度和依赖于视图的RGB颜色。对于多视图立体视觉，由于CNN和MLP无法处理任意数量的输入视图，因此首先独立处理每个视图坐标系中的坐标和相应特征，并获得每个视图的图像条件中间表征。接下来，用基于辅助池化的模型聚合这些NeRF网络内的视图中间表征。在3D理解任务中，多视图提供场景的附加信息。

论文【11】提出一个编码器-解码器Transformer框架TransNeRF，表征神经辐射场场景。TransNeRF可以探索多视图之间的深层关系，并通过单个基于Transformer的NeRF注意机制将多视图信息聚合到基于坐标的场景表征中。此外，TransNeRF考虑光线投射空间和周视空间的相应信息来学习场

景中形状和外观的局部几何一致性。

如图13所示，TransNeRF在一个目标视线（target viewing ray）渲染所查询的3D点，TransNeRF包括：1）在周视空间中，密度-视图解码器（Density-ViewDecoder）和颜色-视图解码器（Color-ViewDecoder）将源视图和查询空间信息*((x,y,z),d)*融合到3D查询点的潜密度和颜色表征中；2） 在光线投射空间中，用密度光线解码器（Density-RayDecoder）和颜色光线解码器（Color-RayDecoder），考虑沿目标视图光线的相邻点来增强查询密度和颜色表征。最后，从TransNeRF获得在目标视线上查询3D点的体密度和方向颜色。



论文【12】提出一种稀疏输入的可泛化NVS方法，称为FWD，实时提供高质量的图像合成。通过显式深度和可差分渲染，FWD实现130-1000倍的速度和更好的感知质量。如果在训练或推理期间有传感器深度的无缝集成，可提高图像质量同时保持实时速度。

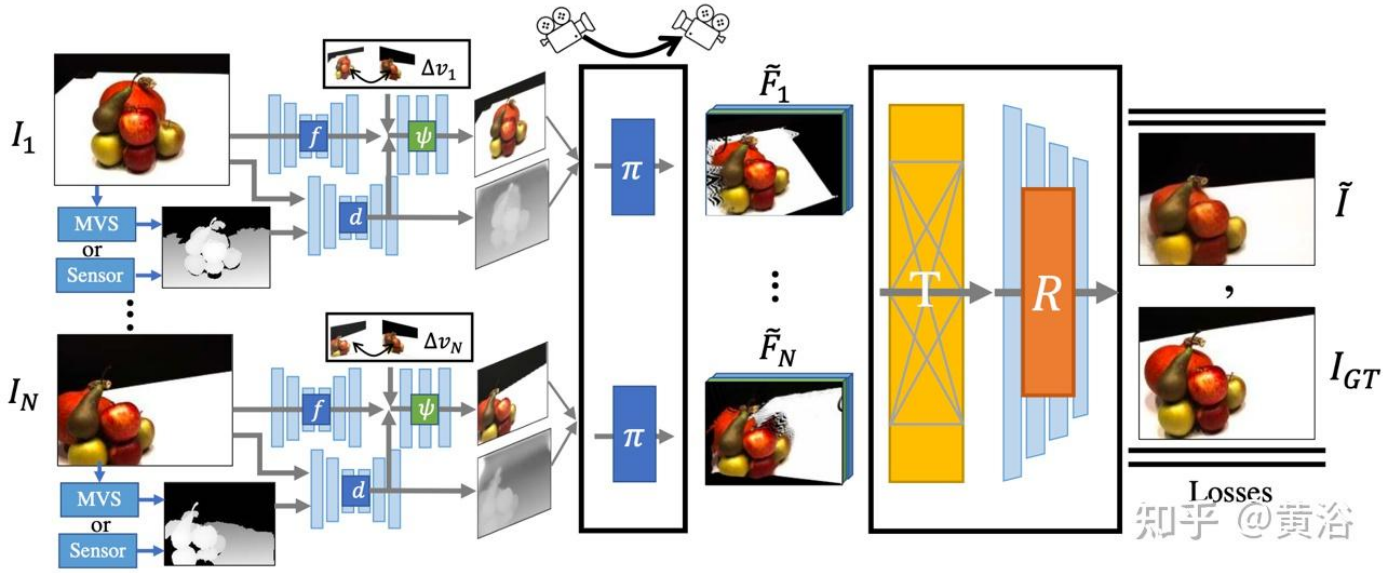
其关键见解是，显式表征每个输入像素的深度允许用可微分点云渲染器对每个输入视图应用forward warping。这避免了NeRF类方法昂贵的体采样，实现了实时速度，同时保持了高图像质量。

SynSin【1】为单图像新视图合成（NVS）使用可微分点云渲染器。论文【12】将SynSin扩展到多输入，并探索了融合多视图信息的有效方法。

FWD估计每个输入视图的深度，构建潜特征的点云，然后通过点云渲染器合成新视图。为了缓解来自不同视点观测之间的不一致问题，将视点相关的特征MLP引入到点云中，对视点相关结果进行建模。另外一种基于Transformer的融合模块，有效地组合来自多输入的特征。一个细化模块，可以修复（inpaint）缺失区域并进一步提高合成质量。整个模型经过端到端训练，最小化光度和感知损失、学习能优化合成质量的深度和特征。

如图14为FWD的概览：给定一组稀疏图像，用特征网络*f*（基于BigGAN架构）、视图相关特征MLP ψ和深度网络*d*为每个图像 I_i构建点云（包括视图的几何和语义信息）P_i；除图像外，*d*将MVS（基于PatchmatchNet）估计的深度或传感器深度作为输入，并回归细化的深度；基于图像特征F_i和相对视图变化 Δv（基于归一化视角方向v_i和v_t，即从点到输入视图*i*和目标视图*t*的中心），通过*f*和ψ回归逐像素特征F_i'；采用可微分点云渲染器 π（splatting）将点云投影和渲染到目标视

图，即 $\tilde{F}_i$；渲染前不是直接聚合视图点云，而是Transformer *T*融合来自任意数量输入的渲染结果，并应用细化模块*R*解码生成最终图像结果，即以语义和几何的方式修复输入看不见的区域，纠正由不准确深度引起的局部误差，并基于特征图所包含的语义提高感知质量；模型训练使用光度损失和内容损失。

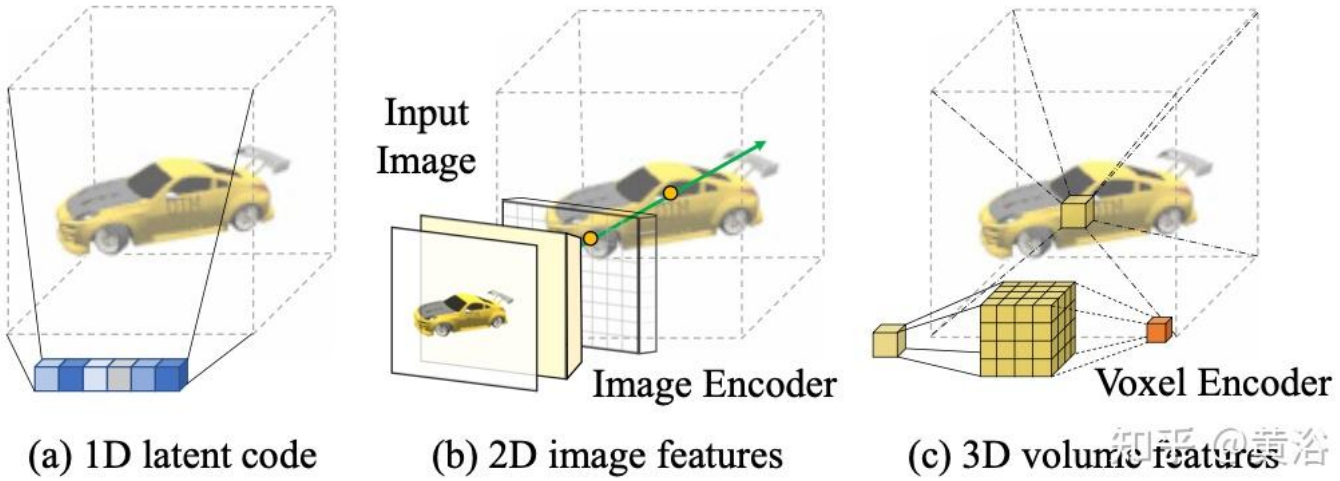


现有用局部图像特征重建3D目标的方法，在查询3D点上投影输入图像特征来预测颜色和密度，从而推断3D形状和外观。这些图像条件模型可以很好地渲染接近输入视角的目标视角图。然而，当目标视角过多移动时，这种方法会导致输入视图的显著遮挡，渲染质量急剧下降，呈现模糊预测。

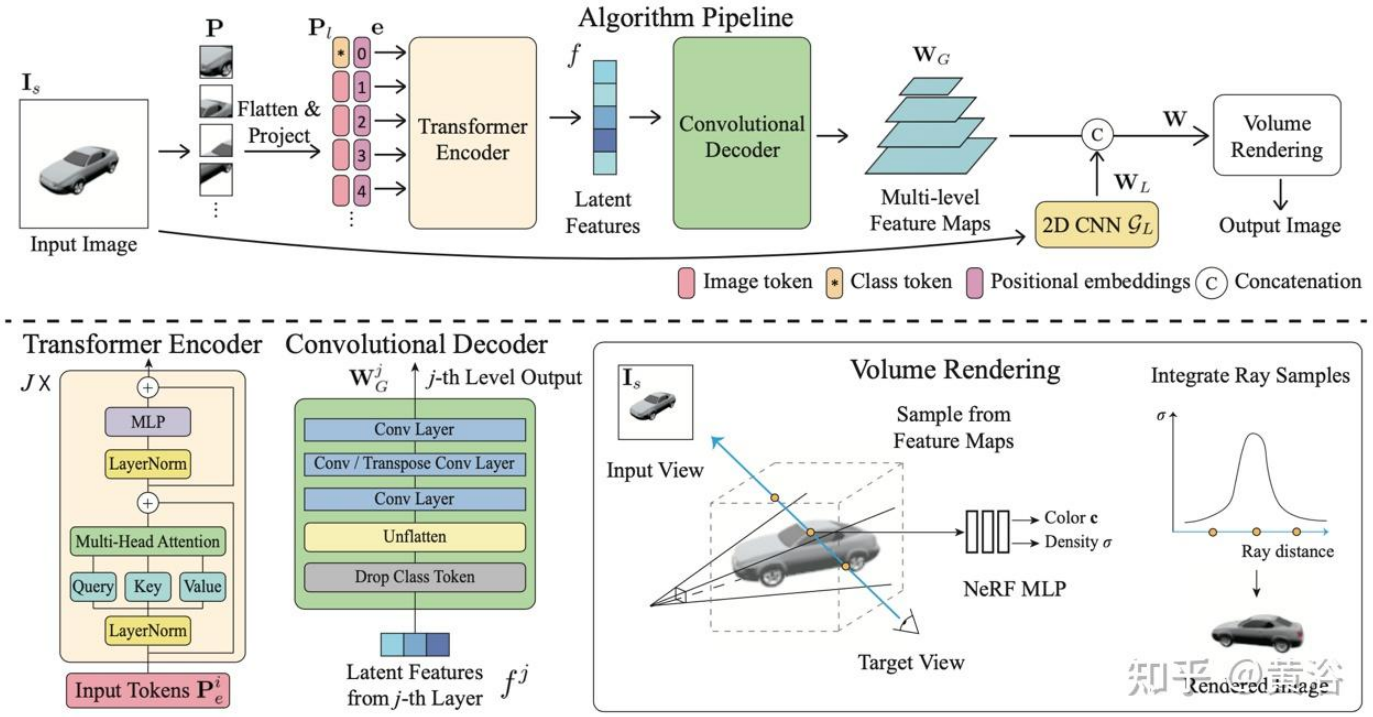
为了解决上面的问题，论文【13】提出一种方法，利用全局和局部特征形成一个压缩的3D表征。全局特征从视觉Transformer中学习，而局部特征从2D卷积网络中提取。为了合成一个新视图，训练了一个MLP网络，根据学习的3D表

征实现体渲染。这种表征能够重建未见过的区域，无需如对称或规范坐标系的强制约束。

给定摄像头s处的单个图像Is，任务是在摄像头t处合成新视图It。如果一个3D点x在源图像中可见，可以直接用其颜色Is(π(x))，其中π代表在源视图进行投影，表示该点在一个新视图可见。如果x被遮挡，就求助于在投影π(x)颜色以外的信息。如图15所示，得到此类信息有三种可能的解决方案：（a）一般NeRF 基于1D潜代码的方法，在1D向量中编码3D目标信息，由于不同3D点共享同一个代码，归纳偏差被限制；（b） 基于2D图像的方法，从逐像素图像特征重建任何3D点，这样的表征鼓励可见区域更好的渲染质量，计算也更有效，但是对未见区域渲染变得模糊；（c） 基于3D体素的方法将3-D目标视为体素的一个集合，并应用3-D卷积生成颜色RGB和密度向量σ，这样渲染较快，也充分利用3D先验去渲染未见的几何，但是由于体素大小和有限的感受野原因限制了渲染分辨率。



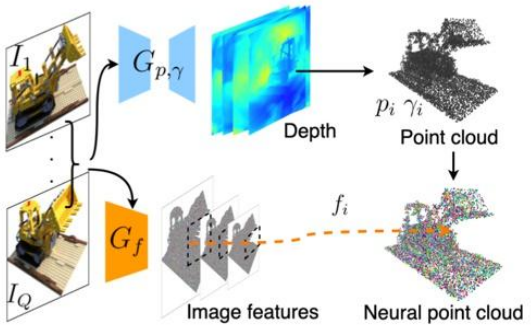
如图6是全局-局部混合渲染方法【13】的总览图：首先将输入图像划分为N=8×8个图像块***P***；每个图像块扁平化并线性投影到图像标记（token）P1；transformer编码器将图像标记和可学习位置嵌入e作为输入，提取全局信息作为一组潜特征f；然后，用卷积解码器将潜特征解码为多级特征图 W_G；除了全局特征，用另一个2D CNN 模型获取局部图像特征；最后，用NeRF MLP模型对体渲染的特征进行采样。



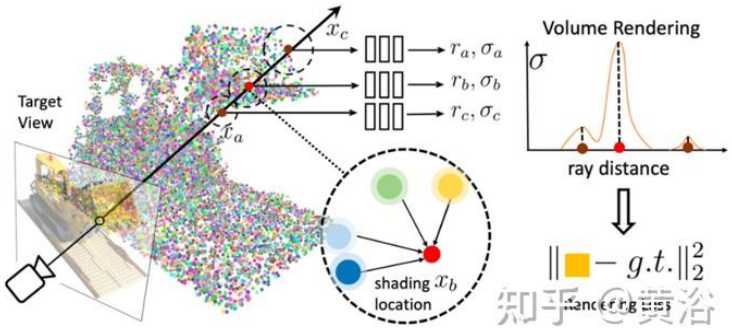
论文【14】提出Point-NeRF，结合NeRF和MVS这两种方法的优点，用神经3D点云以及相关的神经特征对辐射场建模。在基于光线行进的渲染流水线中聚集场景表面附近的神经点特征，可以有效地渲染Point-NeRF。此外，一个预训练的深度网络直接推断可初始化Point-NeRF，生成一个神经点云；该点云可进行微调，超过NeRF的视觉质量，训练时间快30倍。

Point-NeRF与其他3D重建方法相结合，并采用生长和修剪机制，即在高体密度区域生长和在低体密度修剪，对重建点云数据进行优化。

Point-NeRF概览如图17所示：（a） 从多视图图像中，Point-NeRF用基于成本体的3D CNN 为每个视图生成深度，并通过2D CNN 从输入图像中提取2D特征；聚集深度图后，获得基于点的辐射场，其中每个点具有空间位置、置信度和未投影的图像特征；（b） 为合成一个新视图，进行可微分光线行进，并只在神经点云附近计算明暗；在每个明暗位置，Point-NeRF聚集来自其K个神经点邻居的特征，并计算辐射率和体密度，然后用体密度累积求和辐射度。整个过程端到端可训练，基于点的辐射场可以通过渲染损失进行优化。



(a) Neural Point Generation.

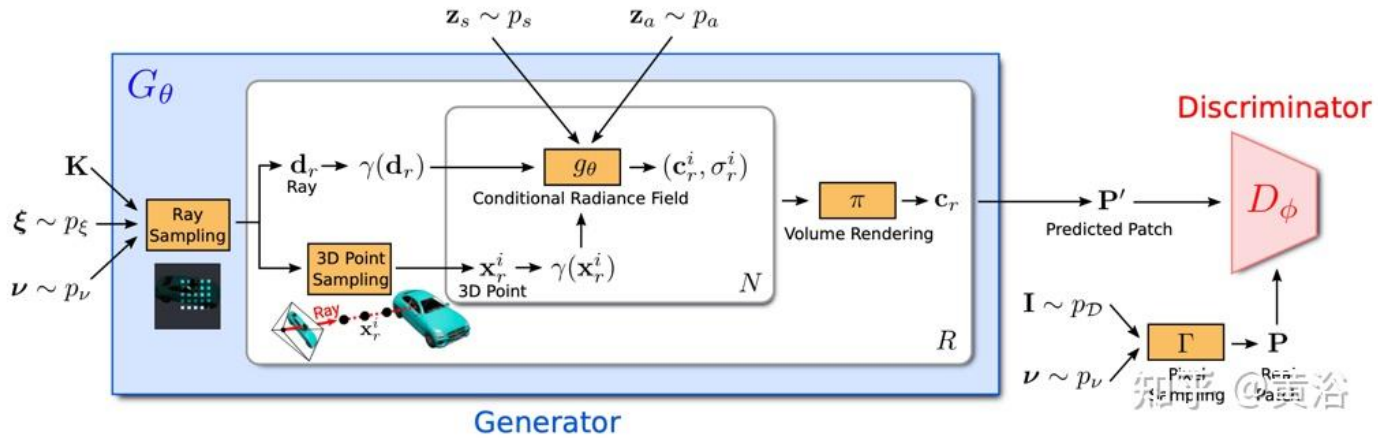


(b) Point-NeRF Representation with Volume Rendering.

GRAF（Generative Radiance Field） 【18】是一种辐射场的生成模型，通过引入基于多尺度patch的鉴别器，实现高分辨率3D-觉察图像的合成，同时模型的训练仅需要未知姿态摄像头拍摄的2D图像。

目标是学习一个模型，通过对未经处理的图像进行训练来合成新的场景。更具体地说，利用一个对抗性框架来训练一个辐射场的生成模型（GRAF）。

图18显示了GRAF模型的概述：生成器采用摄像机矩阵***K*** 、摄像机姿态**ξ** 、2D采样模式***v*** 和形状/外观代码作为输入并预测一个图像patch ***P′*** ；鉴别器将合成的patch ***P′*** 与从真实图像***I*** 中提取的patch ***P*** 进行比较；在推理时，为每个图像像素预测一个颜色值；然而在训练时间这个操作太贵，因此预测一个大小为***K×K*** 像素的固定patch，其随机缩放和旋转，为整个辐射场提供梯度。

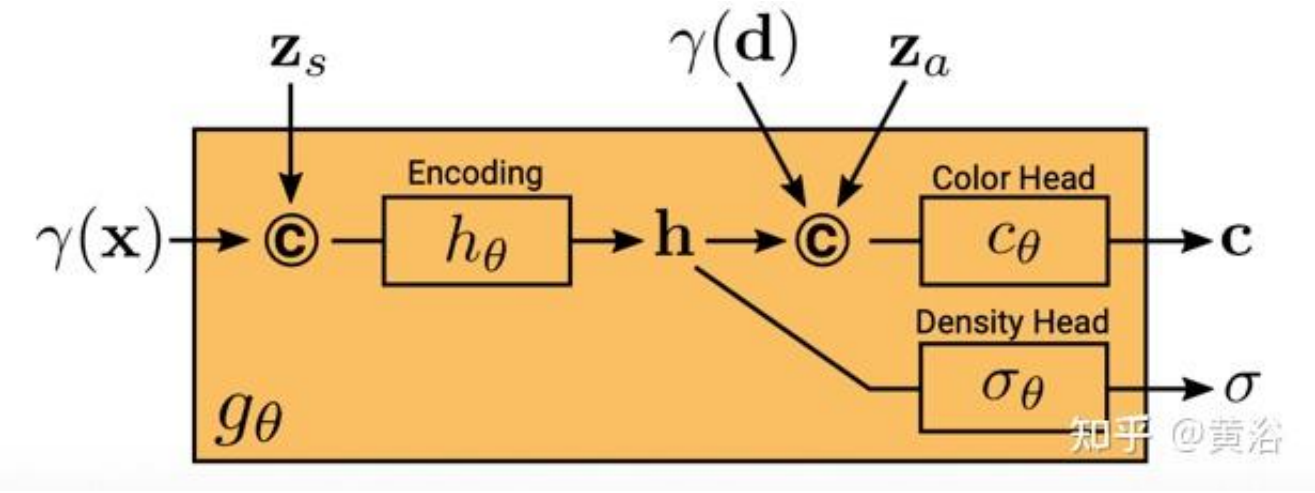


决定要生成虚拟K×K patch 的中心和尺度s。随机patch中心来自一个图像域Ω的均匀分布，而patch尺度s来自一个均匀分布，其中，其中W和H表示目标图像的宽度和高度。形状和外观变量的采样分别来自形状和外观分布和。在实验中，和都使用标准高斯分布。

辐射场由深度全连接的神经网络表示，其中参数θ映射3D位置***x*** 的位置编码和观察方向***d*** 到RGB颜色值***c*** 和体密度σ：

$$g_\theta:\ (\gamma(\boldsymbol{x}),\gamma(\boldsymbol{d}),\boldsymbol{z}_s,\boldsymbol{z}_a)\rightarrow(\boldsymbol{c},\sigma)$$

这里 g_θ取决于两个附加潜代码：一个是形状代码z_s决定目标形状，一个表观代码z_a决定外观。这里称g_θ为条件辐射场，其结构如图19所示：首先根据***x*** 的位置编码和形状代码计算形状编码***h*** ；密度头σ_θ将此编码转换为体密度σ；为预测3D位置***x*** 处的颜色***c*** ，将***h*** 与***d*** 的位置编码以及表观代码z_a连接起来，并将结果向量传递给颜色头c_θ；独立于视点***d*** 和外观代码计算σ，鼓励多视图一致性，同时形状与外观进行分离；这个鼓励网络用两个潜代码分别对形状和外观建模，并允许在推理过程中做分别处理。

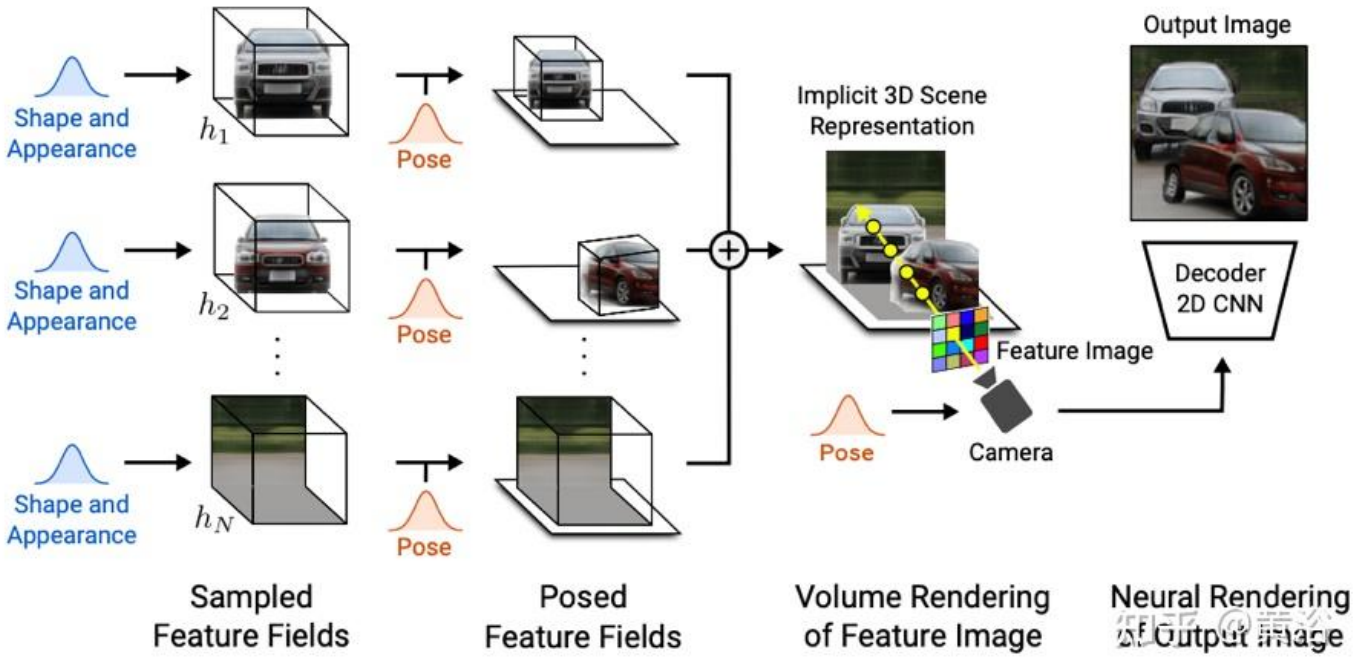


鉴别器实现为一个卷积神经网络，将预测的patch ***P*** '与从数据分布 p_D真实图像***I*** 中提取的patch ***P*** 进行比较。为了从真实图像***I*** 提取*K*×*K*patch，首先从用于提取上述生成器patch的同一分布p_v中提取v=(u,s)；然后，通过双线性插值在2D图像坐标P(u,s)处查询***I*** ，采样真实patch ***P*** 。用 $\Gamma(I,v)$表示这种双线

性采样操作。

实验发现一个有共享权重的单鉴别器足以用于所有patch，即使这些patch在不同尺度随机位置采样。注: 尺度决定patch的感受野。因此，为了促进训练，从更大的接受野patch开始去捕捉全局上下文。然后，逐步采样具有较小感受野的patch细化局部细节。

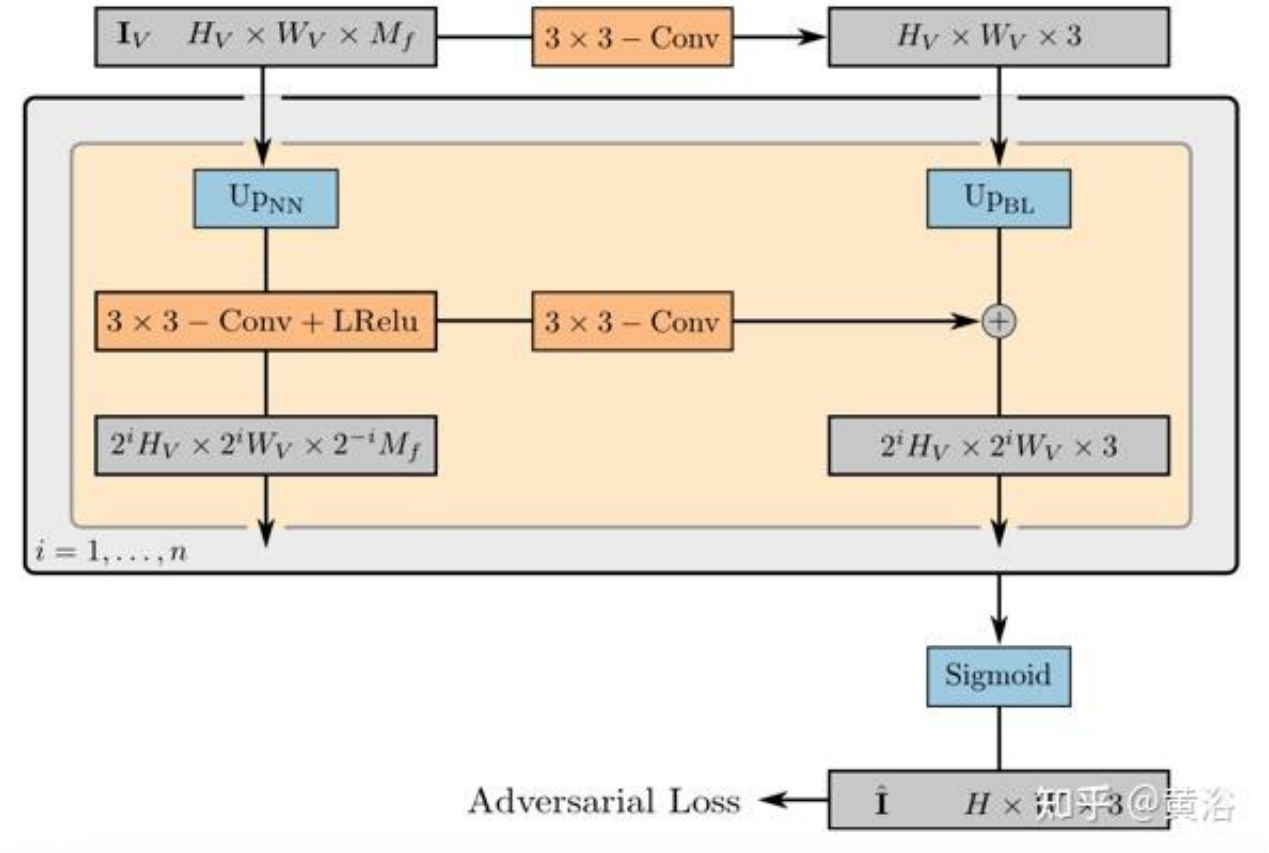
GIRAFFE 【19】用于在原始非结构化图像进行训练时以可控和真实感的方式生成场景。主要贡献有两个方面：1) 将组合3D场景表征直接纳入生成模型，实现更可控的图像合成。2) 将这种明确的3D表征与一个神经渲染流水线相结合，实现更快的推理和更逼真的图像。为此，场景表征为**组合生成神经特征场**，如图20所示：对于一个随机采样的摄像头，基于单独特征场对场景的一个特征图像进行体渲染；2D神经渲染网络将特征图像转换为RGB图像；训练时只采用原始图像，在测试时能够控制图像形成过程，包括摄像头姿势、目标姿势以及目标的形状和外观；此外，该模型扩大到训练数据范围之外，例如，可以合成包含比训练图像中更多目标的场景。



将场景体渲染为分辨率相对较低的特征图像，可节省时间和计算。神经渲染器处理这些特征图像并输出最终渲染。通过这种方式，该方法可以获得高质量的图像并尺度化到真实场景。当在原始非结构化图像集合上进行训练时，这个方法允许单目标和多目标场景的可控图像合成。

场景组合时，要考虑两种情况：N固定和N变化（其中最后一个是背景）。在实践中，像目标那样，背景用相同的表征法，不同的是横跨整个场景把尺度和平移参数固定，并以场景空间原点为中心。

2D渲染算子的权重把特征图像映射到最后合成图像，可以参数化为一个带泄漏ReLU激活的2D CNN，和3x 3卷积和最近邻域上采样结合可增加空域分辨率。最后一层应用sigmoid操作，得到最后的图像预测。其示意图如图21所示。



鉴别器也是一个带泄漏ReLU激活的CNN。

参考文献

1. O Wiles, G Gkioxari, R Szeliski, and J Johnson. “Synsin: End-to-end view synthesis from a single image”. IEEE/CVF CVPR, 2020.
2. A Chen, Xu, FuZhao, X Zhang, F Xiang, J Yu, and H Su. “Mvsnerf: Fast general- izable radiance field reconstruction from multi-view stereo”. *arXiv 2103.15595*, 2021.

3. J Chibane，A Bansal，V Lazova，G Pons-Moll，“Stereo Radiance Fields (SRF): Learning View Synthesis for Sparse Views of Novel Scenes”， arXiv2104.06935, 2021
4. A Jain, M Tancik, and P Abbeel. “Putting nerf on a diet: Semantically consistent few-shot view synthesis”（DietNeRF）. arXiv 2104.00677, 2021.
5. J-Y Zhu，K Deng, A Liu and D Ramanan. “Depth-supervised NeRF: Fewer views and faster training for free”（DS-NeRF）. arXiv 2107.02791, 2021
6. A Yu, V Ye, M Tancik, and A Kanazawa. “pixelNeRF: Neural radiance fields from one or few images”. IEEE *CVPR*, 2021.
7. Q. Wang, Z. Wang, K. Genova, P. P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser, “IBRnet: Learning multi-view image-based rendering,” *IEEE/CVF CVPR*, 2021
8. A Trevithick and B Yang. “GRF: Learning a general radiance field for 3d representation and rendering”. IEEE/CVF CVPR, 2021.

9. M Niemeyer, J T. Barron, B Mildenhall, M S. M. Sajjadi, A Geiger, N Radwan, “RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs“, arXiv 2112.00724, 2021
0. ”J Zhang, Y Zhang, H Fu, X Zhou, B Cai, JHuang, R Jia, B Zhao, X Tang, “Ray Priors through Reprojection: Improving Neural Radiance Fields for Novel View Extrapolation”, arXiv 2205.05922, 2022
1. D Wang, X Cui, S Salcudean, and Z. J Wang, “Generalizable Neural Radiance Fields for Novel View Synthesis with Transformer”, arXiv 2206.05375, 2022
2. A Cao, C Rockwell, J Johnson, “FWD: Real-time Novel View Synthesis with Forward Warping and Depth ”, arXiv 2206.08355, 2022
3. K-E Lin, L Y-C, W-S Lai, R Ramamoorthi, Y-C Shih, T-Y Lin, “Vision Transformer for NeRF-Based View Synthesis from a Single Input Image ”, arXiv 2207.05736, 2022
4. Q Xu, Z Xu, J Philip, S Bi, Z Shu, K Sunkavalli, U Neumann, “Point-NeRF: Point-based Neural Radiance Fields”, arXiv 2201.08845, 2022

5. R Tucker and N Snavely，“Single-view view synthesis with multiplane images”（MPI）. IEEE/CVF CVPR, 2020.
6. J Li, Z Feng, Q She, H Ding, C Wang, G H Lee，“MINE: Towards Continuous Depth MPI with NeRF for Novel View Synthesis ”，arXiv 2103.14910，2021
7. B Roessle，J T. Barron，B Mildenhall，P P. Srinivasan，M Niesner，“Dense Depth Priors for Neural Radiance Fields from Sparse Input Views”, arXiv 2112.03288，2021
8. K Schwarz, Y Liao, M Niemeyer and A Geiger, “GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis”, NeurIPS 2020
9. M Niemeyer and A Geiger “GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields”, IEEE CVPR 2021