基于残差颜色学习的新视图合成

韩磊，钟大伟，李林，郑凯，方陆，*IEEE高级会员*

***摘要:场景表示网络(scene* Representation Networks, SRN)在近年来的研究中已被证明是一种强大的新视图合成工具。它们使用全连接网络学习从空间点的世界坐标到辐射颜色和场景密度的映射函数。然而，场景纹理在实践中包含复杂的高频细节，这些细节很难被参数有限的网络记忆，导致在渲染新视图时出现令人不安的模糊效果。本文建议学习“残差颜色”而不是“辐射颜色”来进行新视图合成，即表面颜色和参考颜色之间的残差。这里的参考颜色是基于空间颜色先验计算的，这些先验是从输入视图观察中提取的。这样一种策略的美妙之处在于，对于大多数空间点来说，辐射颜色和参考颜色之间的残差接近于零，因此更容易学习。提出了一种利用SRN学习残差颜色的视图合成系统。在公开数据集上的实验表明，所提出的方法在保留高分辨率细节方面取得了有竞争力的性能，导致了比目前最先进的技术在视觉上更令人愉快的结果。**

索引术语-神经渲染，图像生成，视图合成，立体视觉。

I. 介绍

新视图合成，作为虚拟现实应用的基本技术，旨在从场景的给定观察样本中创建新视图。最近的工作，如GoogleJump [2]， DeepView[6]等，通过使用同步结构化相机阵列作为捕获设备，已经显示出重大进展。然而，从稀疏视图输入进行高质量新视图合成仍然是一项具有挑战性的任务。现有方法试图通过重建场景[7]，[10]的显式几何模型或采用概率深度表示[27]，[37]来解决该问题。通常，基于模型的方法在很少的输入视图下享有更高的自由度，但需要高分辨率和精确的3D模型。此外，它不能反映这种变化.3154242的不同视角的光。另一方面，基于概率的深度方法将场景几何建模为概率分布，而不是显式的深度表面。例如，StereoMagnify[37]采用多平面图像进行场景表示，并基于alpha合成呈现新颖视图;NeRF[24]使用隐式场景表示网络将场景参数化为一个亮度场，并应用体绘制进行新颖视图合成。

在最新的研究中，NeRF[24]通过使用完全连接的网络来表示复杂场景的底层连续体辐射场，实现了卓越的性能。该网络可以直接在一组稀疏的输入2D图片上进行训练，而无需额外的3D监督。得益于体积场景表示，NeRF为自由移动的相机生成连续的新颖视图合成。不幸的是，由于神经网络过度拟合低频信息[28]的固有性质，即使使用位置编码方案，合成的图像也丢失了高频纹理细节，这导致了令人不安的模糊效果。

我们认为当前隐含场景表示网络简单的编码空间坐标表示的每一个点而忽视点可能自己differ-ent特点back-projected到输入时的观点。具体来说，不同视角下的反投影观测值(记为空间颜色先验)对于朗伯曲面上的点是一致的，而对于非曲面点则有显著变化。因此，空间颜色先验和每个点的实际辐射亮度颜色之间存在着很强的联系。

基于这一观察，本文提出了一种用于新视图合成的残差颜色学习框架。具体而言，我们将每个点的空间颜色先验作为参考颜色，并使用场景表示网络(如NeRF[24])来回归表面颜色与参考颜色之间的残差。图1显示了我们渲染结果的分解。注意，对于大多数空间点，残差都是小值或接近于零。因此，它们比之前直接强制网络记忆错综复杂的纹理细节的方法更容易学习。所提出方案为新视图合成保留了更清晰的细节，比最先进的方法带来了更令人满意的视觉结果。值得注意的是，对于复杂场景，以前的方法如NeRF[24]存在模糊的伪影，而我们的方法由于残差学习方案而取得了显著的改进。总结了如下的技术贡献。

· 空间颜色先验:鉴于多视角观测传达了辐射亮度颜色的先验信息，

稿件收到于2021年3月24日;修订于2021年10月30日和2022年1月9日;审定2022年2月8日。出版日期2022年3月2日;当前版本发布日期2022年3月11日。本工作得到中国自然科学基金(NSFC)合同62125106、合同61860206003、合同62088102的部分支持;部分由中国科学技术部按合同2021ZD0109901承担;部分由深圳市科学技术研发基金项目JCYJ20180507183706645资助;部分由北京国家信息科学技术研究中心(BNRist)资助，项目编号BNR2020RC01002。协调审稿并批准发表的副主编是陈振忠博士。*(韩磊、钟大伟对本文贡献相同。)(通讯作者:方陆。)*

韩磊，李林，郑凯，海思半导体，中国上海518129。

钟大伟、方陆在清华大学电子工程系、清华-伯克利深圳研究所工作，北京100190;北京国家信息科学技术研究中心(BNRist)，北京100084 (e-mail: fanglu@tsinghua.edu.cn)。

数字对象标识符10.1109/ tip .2022

图1所示。对于3D空间中的每个点，我们根据多视角观察计算其参考颜色。由场景表示网络记忆的残差颜色以及密度。参考彩色图像(a)和残差彩色图像(b)是基于采样点的体绘制而组成的，并组合用于新视图合成。注意，参考彩色图像包含大部分高频纹理信息，而SRN只需要表示由低频信息组成的残差颜色和几何形状(密度场)。如(d)和(e)所示，对于新视图合成，所提出的方法比最先进的方法NeRF[24]生成更清晰的细节。

所提出的学习框架配备了基于输入视图观察的空间颜色先验，这是隐式场景表示网络的补充信息，该网络只是将点的世界坐标映射到局部场景属性。

· 残差颜色学习:通过将提出的空间颜色先验作为参考，我们提出了一种残差颜色学习框架，以回归表面颜色和参考之间的残差。对于大多数空间点来说，残差接近于零，因此比之前直接回归表面颜色的工作更容易学习。在神经渲染中提出的残差学习框架简单而有效，可以很容易地与其他隐式场景表示方法相结合。

II. 相关工作

真实感渲染旨在基于有限的观察生成任意的新视图，主要分为两种不同的管道:基于纹理的渲染和基于图像的渲染。基于纹理的渲染遵循经典的渲染管道，构建显式的3D模型，并基于光线跟踪获得渲染图像。而基于图像的渲染使用软3D表示，如概率深度或神经网络，用于没有显式3D模型的隐场景表示。下面，我们将分别介绍基于纹理的渲染和基于图像的渲染的当前进展。

A. Texture-Based呈现

基于纹理的绘制旨在重建环境的精确彩色三维模型，以实现新视点绘制。[7]，[10]利用多视角观测和极线几何的密集匹配来重建3D模型。Elastic-Fusion[33]使用帧到模型配准和基于窗口的surf融合。[36]采用基于空间哈希[17]的体积融合和TSDF融合[18]实现实时重建。随着机器学习的发展，神经网络也被用于预测显式3D模型。[16]，[32]将2D特征投影到3D体素网格，并使用3D卷积得到体素模型。[20]使用可微的基于点的渲染器来获得3D模型。点的坐标和颜色是学习参数。[8]使用多层感知器将点云完成成网格模型。[13]训练一个基于块的条件判别器来指导纹理优化，使其能够容忍错误对齐。其性能受限于现有3D模型的质量。

在显式3D模型的帮助下，基于纹理的渲染具有良好的效率和可编辑性。然而，在重建的模型中很难避免失真、空洞和模糊的部分，尤其是对于凌乱的场景。生成模型的不足会在渲染图像时带来伪影和模糊的细节。

B. 基于图像的渲染

与基于纹理的渲染不同，基于图像的渲染无需明确的3D模型即可生成新颖的视图。[4]，[12]通过转换采样图像生成新视图。采样图像扭曲成一个新颖的观点基于相机姿态估计的估计。[5]，[14]使用贝叶斯估计来获得新视图中每个像素的颜色值。神经网络被广泛应用于隐式场景表示。它在记忆场景方面显示出巨大的潜力，包括几何和纹理。几何由于其低频性质，可以很容易地用神经网络来表示，而高频纹理细节则很难被神经网络记忆。[6]、[37]在不同层生成不同透明度的多平面图像，通过对多平面图像的整合可以得到新的视图。[23]通过变换的相邻多平面图像的加权组合生成新视图，这些图像由相应的透明度调制。[25]推导出用于TSDF值预测的可微体绘制。[31]将世界坐标映射到局部场景属性的特征表示，并使用场景表示网络来预测不同类型场景的特殊网络。[22]使用编码器根据多视图图像产生一个潜代码z，然后将其解码为一个volume，该volume为每个体素提供颜色和透明度值。NPBG[1]将一组RGB视图和一个点云作为输入。对每个点拟合一个神经描述子，之后可以渲染一个场景的新视图。FVS[30]通过多视图立体计算输入图像的3D代理几何。给定一个目标视图，根据投影深度将附近的源图像映射到目标视图中，然后使用循环卷积网络将映射后的图像进行混合。这两种方法都需要高质量的3D几何图形作为输入。渲染性能很大程度上受点云或重建的3D几何的质量影响。如果用于映射的3D模型遗漏了场景的大部分或有粗糙的异常值，管道将产生可见的伪影。NeRF[24]通过多层感知器表示一个场景，并通过体绘制对其进行训练。使用位置编码和分层采样来提高渲染性能。

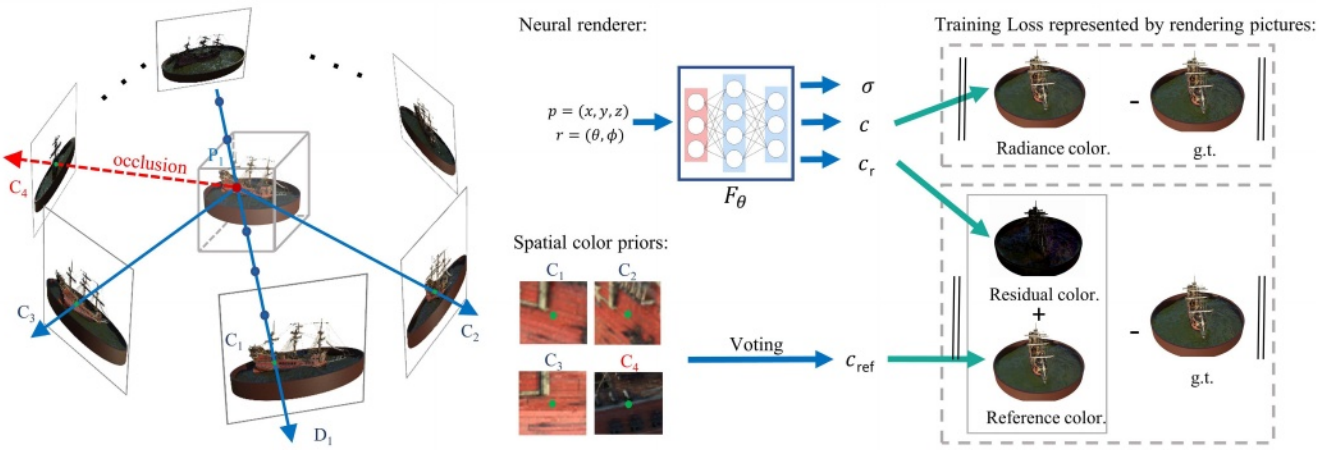

图2所示。所提出的残差颜色学习方案概述。对于每个空间点，我们从输入的多视图观察中计算其参考颜色*cref*，并使用场景表示网络预测密度σ,辐射度颜色*c*和残差颜色c_r。空间颜色先验是投影像素(例如*P1*的*C1, C2, C3*和*C4*，它们是所呈现的图像块的中心像素)。图像块用于过滤掉被遮挡的像素。点*P1*的参考颜色*cref*是通过对点的反向投影像素进行投票来估计的。对于给定视点的新视图合成，基于预测的密度σ，通过沿着所有像素射线整合空间点c_r+c_{ref}来应用体*绘制*。集成辐射亮度颜色*c*来预测遮挡检测的粗略图像(在本例中去除*C4*)，以更好地进行参考颜色预测。在训练阶段，输入视图被采样为真实值，*F(θ)*使用辐射色和残差色的渲染损失进行训练。

隐场景表示在真实感绘制中显示出巨大的潜力，但仍然是一项具有挑战性的任务。NeRF[24]使用多层感知器和体积渲染进行隐式表示。它实现了显着的渲染性能改进，并且有许多方法可以提高NeRF。NSVF和我们的方法都旨在提高不同角度的新视图合成的渲染质量。NSVF利用三维空间中表面稀疏的先验性，只需要处理经过表面的体素，并使用局部参数来提高场景表示网络的能力。不同的是，我们的方法提出了空间颜色先验，通过从投影像素计算参考颜色来降低高频纹理细节的学习难度。Nerfies[26]引入变形代码来处理动态场景，并使用外观代码来处理光线变化。KiloNeRF[29]利用数千个微小MLP来取代原来单一的大型MLP进行加速。我们的方法是对此类方法的补充。为了提高保持高频纹理的能力，本文基于多视角观察提出了新的基于残差的多视角先验。利用所提出的空间颜色先验，为高质量的隐场景表示引入了残差学习方案。

III. 方法

该方法将一组稀疏的视图作为输入，旨在在给定的视点上渲染新视图。整体框架如图2所示。我们根据输入的多视角观察提出“空间颜色先验”，而被遮挡的像素则通过提出的patch特征滤波器被去除。参考颜色通过投票策略从空间颜色先验中获得。在此基础上，在隐场景表示网络中引入残差颜色学习方案，以降低对高频信息的网络容量要求。

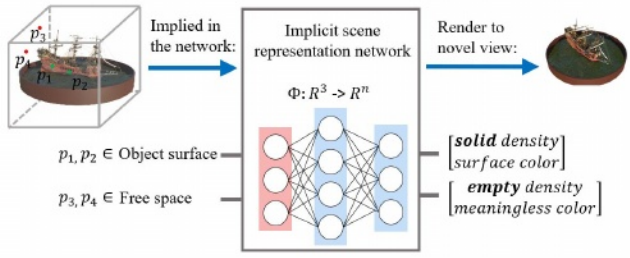

图3所示。与传统的具有显式模型的3D重建不同，隐式场景表示使用函数来拟合场景信息。该函数以点的位置作为输入，并输出其可以以图像形式呈现的空间特征。通过这种方式，采样图像可以用于训练函数，然后可以用于生成新视图。

在接下来的内容中，我们首先介绍了章节III-A中的隐式场景表示，然后详细阐述了章节III-B中的空间颜色先验和章节III-C中的残差颜色学习方案，最后在章节III-D中给出了具体的实现细节。

A. 隐场景表征

[31]采用全连接网络来隐式描述场景。它学习一个函数，该函数将连续的3D坐标映射到这些特征坐标处场景的特征表示。对于不同的目标，特征表示可能被转换为诸如密度[24]或符号距离函数[25](图3)等属性。

代表性的SRN方法NeRF[24]将场景建模为神经辐射场，并应用体积渲染[15]进行新的视图合成。每个空间点由其3D坐标$p = (x, y, z)$和视角方向$d_r = (\theta, \phi)$*表示*，使用全连接网络将其映射到密度(不透明度)σ*和*辐射度颜色*c*。摄像机光线*r的期望颜色C(r)*可以是

由经典的体绘制技术绘制，如式1所示。

$$\bar{C}(r) = \int_{t_n}^{t_f} \exp\Big(-\int_{t_n}^{t} \sigma(s)ds\Big)\sigma(t)c(t, d_r)dt, \quad (1)$$

其中t_n和t_f分别是r的近界和远界，dt是相机光线之间的距离。d_r表示r的视点方向，“exp”为指数函数。基于体绘制[15]，式1的连续积分可以用数值正交代替:

$$\begin{aligned} T_i &= \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j), \\ w_i &= T_i\big(1 - \exp(-\sigma_i \delta_i)\big), \\ \bar{C}(r) &= \sum_{i=1}^{N} w_i c_i. \end{aligned} \quad (2)$$

σ_i,c_i用一个全连接的网络$F\theta(p_i, d_r)$表示，分别表示第i个采样点的颜色和密度。δ_i表示两个采样点之间的距离。$C(r)^-$是根据权重w_i将射线中的所有采样点相加计算出来的。这里$F\theta(p_i, d_r)$可以通过最小化渲染视图C(r)$^-$和观察视图$C(r)$之间的差异来从给定的稀疏输入视图中学习:

$$L = \sum_{r \in R} \|\bar{C}(r) - C(r)\|, \quad (3)$$

其中R是所有相机射线的集合。它的个数等于所有图像像素的个数。

B. 空间色彩先验

回想一下,现场几何和纹理信息隐含在颜色的一致性的多视图的观察,在此基础上,我们提出“空间颜色先知先觉”和剩余颜色学习计划来减少高频信息的网络容量需求。

首先将空间点投影到观测图像上，得到其投影直方图;训练图像记为I ={I_i, I∈N}，对应的相机位姿记为$\boldsymbol{H}$ ={H_i, I∈N}。我们计算当前相机姿态H_c与$\boldsymbol{H}$**之间的距离**，并从训练图像$\boldsymbol{I}$中选择M个最近的图像，局部图像为Ilocal = {Ilocali, I∈M}。然后基于多视图几何[9]计算反投影像素:

$$u_i = K H_i H_c^{-1} p, \quad i \in M, \quad (4)$$

其中K是相机的内蕴。$\boldsymbol{u}$ ={u_i, i∈M}是点p在局部图像$\boldsymbol{I}_{local}$中的投影像素，点p的投影直方图被定义为***u*的统计直方图**。

采样点离物体表面的远近会导致投影直方图的特征不同。图4说明了非表面点和表面点两种情况下的投影直方图。对于非表面点，不同视角的观察结果是不相关的，从其散乱投影直方图可以看出。

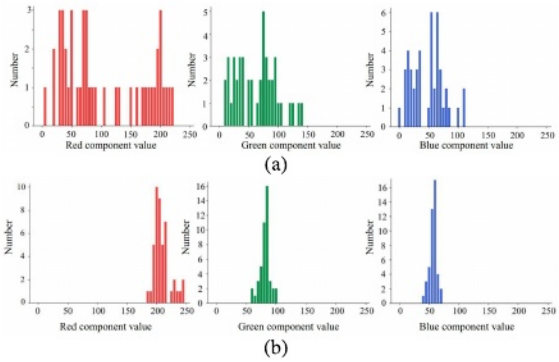

图4所示。空间颜色先验(投影像素的直方图，来自45个投影视图)为(a)一个非表面点和(b)一个表面点的场景的图2。从左到右:观察到的颜色直方图，红色，绿色，蓝色的点，当投影回输入视图。请注意，非表面点(a)的直方图是分布的，而表面上点(b)的直方图是集中的，因此可以根据我们提出的空间颜色先验稳健地估计参考颜色。其他非表面点和表面点有类似的情况。

对于物体表面上的点，不同视角的观测结果是一致的，其投影直方图是集中的。由于投影直方图的颜色一致性隐含了场景几何和纹理信息，因此对于每个空间点，我们基于其投影直方图中的信息提出了“空间颜色先验”。

如果该点在朗伯曲面上，除了被遮挡的像素外，投影像素是相似的。由于被遮挡的像素与其他投影像素无关，因此对于空间颜色先验来说，它们是无意义的噪声。为了处理它，我们采用了一个patch特征过滤器来从投影直方图中去除被遮挡的像素。同一3D点在不同视角下的局部图像块，除了遮挡外，期望是相似的，这适合于遮挡去除。使用半尺寸图片的3 ×3块作为像素特征，因为在相同的局部块大小下，下采样图像具有更大的感受野。将投影像素的patch特征与当前视图进行比较。我们计算l_2范数，删除与当前视图差异大于阈值的像素。提出的patch filter是一种简单但有效的方法来处理周围场景中的多个遮挡。它不需要非常精确，因为残差颜色预测将补偿小偏差。

对于训练，当前视图的patch特征是从训练图像中提取的。对于推理，当前视图的patch特征是从预测的辐射亮度颜色C中提取的(参见第III-C节)。有了patch特征过滤器，被遮挡的像素就可以被移除。之后，我们通过基于剩余投影像素**u的投票策略来计算参考颜色**c_{ref}。虽然我们通过特征滤波去除被遮挡的像素，但一些具有强反射率的投影像素仍然可能影响参考颜色的计算。因此我们计算**u的均值**，然后从均值中删除大于阈值的值。具有强反射率的像素通过投票策略被移除。然后我们通过剩余像素的均值来计算参考颜色。注意，没有

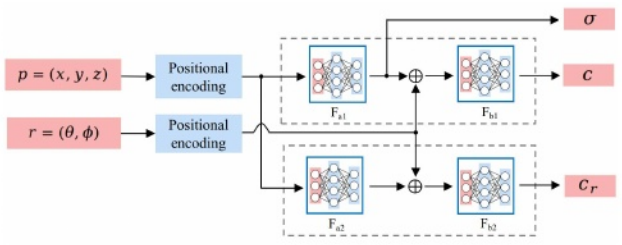

图5所示。网络架构。输入是位置(x, y, z)和观看方向(θ, φ)。输入位置p的位置编码通过8个完全连接的ReLU层传递，每个层有256个通道(Fa1和Fa2)。然后将输出的256特征与输入观看方向r的位置编码相结合，通过4个全连接的ReLU层，每个层有128个通道(Fb1和Fb2)。输出为密度σ,辐亮度颜色c和残差颜色cr。

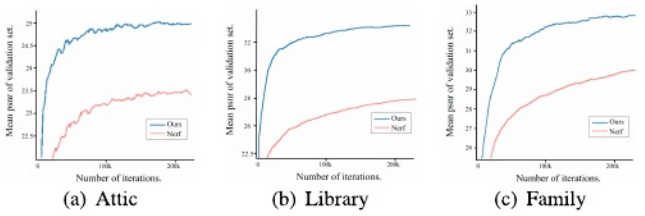

图6所示。NeRF和我们提出的方法在casual 3D数据集[11]的“Attic”(a)和“Library”(b)以及tank and temples数据集[19]的“Family”(c)中的收敛曲线。两种方法都使用相同的训练参数。batchsize为1024，学习率为5−4。我们的方法在相同的迭代次数下实现了更快的收敛。

特征滤波器、残差学习的空间颜色先验已经在大部分区域有了明显的性能提升，但在遮挡时引入了较小的伪影。我们引入特征过滤器来处理遮挡。但是，直接使用特征滤镜会带来更差的结果(图14)，这是因为虽然特征滤镜提供了更准确的参考颜色。它还会使一些非表面点的投影直方图更加集中，从而导致密度预测的准确性降低。特征滤波器必须与Eq. 8的联合训练相结合，以增强密度预测的鲁棒性。然后遮挡中的伪影就会被成功丢弃。

C. 残差颜色学习

通过对空间点的参考颜色计算，本文提出了一种残差颜色学习方案，将空间颜色先验应用于新视图合成。对于每个空间点，我们根据第III-B节所示的空间颜色先验计算其参考颜色c_{ref}，并通过SRN Fθ预测其残差颜色cr。参考颜色和残差颜色被组合为预测颜色c，用于光线R处颜色$\bar{c}R$的体绘制，如下所示:

$$c^i_{com} = c^i_{ref} + c^i_r,$$
$$\bar{C}_R(r) = \sum_{i=1}^{N} T_i\big(1 - \exp(-\sigma_i\delta_i)\big)c^i_{com}. \tag{5}$$

对于朗伯表面上的点来说，不同视图的像素颜色是相似的。通过稳健的参考颜色计算，

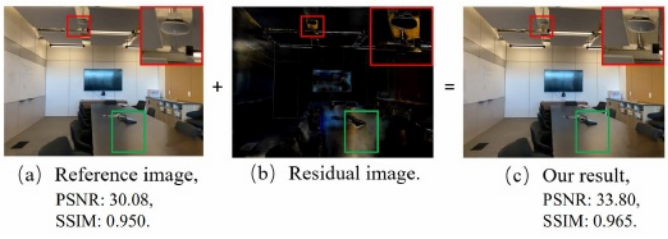

图7所示。渲染结果分解示意图。残差图修正了参考图中因投影像素错误造成的失真(红色块)，并添加了视相关的光阴影(绿色块)。对残差图像进行补救后，渲染结果的PSNR由30.08提高到33.80。

不同视点的残差颜色预测比原始的辐亮度颜色预测小得多。将复杂高频纹理细节的学习任务简化为学习大多数空间点接近0的残差颜色，显著减轻了网络的负担。

$$L = \sum_{r\in R} \|\bar{C}_R(r) - C(r)\| \tag{6}$$

然而，我们也观察到，仅根据Eq. 6所示的残差颜色来学习网络可能会导致过拟合，因为如果非表面点的参考颜色与目标颜色相似，则可能会将其分配给非零密度。辐射颜色和残差颜色可以有各自的密度预测。然而，为了增强引入特征滤波器后密度预测的鲁棒性，我们提出了一种联合训练方案，即通过学习相同密度的残差颜色和辐亮度颜色来利用辐射亮度颜色损失进行密度预测:

$$\bar{C}_W(r) = \sum_{i=1}^{N} T_i\big(1 - \exp(-\sigma_i\delta_i)\big)c_i,$$
$$\bar{C}_R(r) = \sum_{i=1}^{N} T_i\big(1 - \exp(-\sigma_i\delta_i)\big)(c^i_{ref} + c^i_r). \tag{7}$$

δ_i、c_i和$c_r{}^i$是全连接网络的输出$F\theta(p_i,d_r)$。δ_i是密度预测。c_i和$c_r{}^i$分别是辐射色和残差色输出。网络由辐射度图像$\bar{C}_W(r)$和残差图像$\bar{C}_R(r)$的渲染损失共同训练:

$$L = \sum_{r\in R} \|\bar{C}_W(r) - C(r)\| + \sum_{r\in R} \|\bar{C}_R(r) - C(r)\|. \tag{8}$$

所提出的残差颜色学习方案大大减轻了网络的负担。因此，我们提出的方法比NeRF获得了更好的性能，并且迭代次数更少(图6)。

D. 实现细节

根据NeRF[24]，我们为输入视图监督的每个场景训练一个SRN。网络架构如图5所示。在训练步骤中，从训练视图中随机采样像素射线。采用NeRF[24]中提出的分层抽样策略

表我

根据三个指标(PSNR(↑)、ssim(↑)和lpips(↓))对公共数据集进行定量评估。T he分数是所有T测试图像的均值

	Room [23]			Fortress [23]			Drums [24]			Ship [24]		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
SRN [31]	27.29	0.883	0.240	26.63	0.641	0.453	17.18	0.766	0.267	20.60	0.757	0.299
NV [22]		-			-		22.58	0.873	0.214	23.93	0.784	0.276
LLFF [23]	28.42	0.932	0.155	29.40	0.872	0.173	21.13	0.890	0.126	23.22	0.823	0.218
NeRF [24]	32.70	0.948	0.178	**31.16**	0.881	0.171	25.01	0.925	**0.091**	28.65	0.856	0.206
Ours	**32.89**	**0.955**	**0.151**	31.15	**0.905**	**0.144**	**26.06**	**0.934**	0.099	**30.09**	**0.863**	**0.199**

	Library [11]			Attic [11]			Kitchen [11]			Troll [11]		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NeRF [24]	29.02	0.784	0.481	23.64	0.744	0.535	26.13	0.826	0.334	26.04	0.643	0.515
Ours	**33.08**	**0.926**	**0.183**	**25.25**	**0.780**	**0.424**	**27.70**	**0.878**	**0.229**	**26.74**	**0.696**	**0.364**
	Auditorium			Theater			Family [18]			Horse [18]		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NeRF [24]	21.81	0.766	0.334	21.50	0.666	0.425	31.07	0.924	0.126	30.41	0.932	0.144
Ours	**23.58**	**0.834**	**0.210**	**23.38**	**0.691**	**0.323**	**32.71**	**0.953**	**0.069**	**31.08**	**0.948**	**0.104**

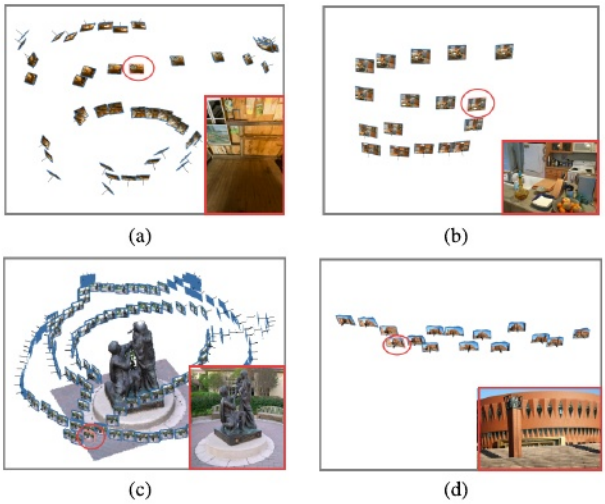

图8所示。4种代表性输入场景的输入场景和推理视点。输入图像根据它们的相机姿态显示。(a)是来自casual 3D[11]数据集的室内周边数据(图8 (a))。(b)是来自LLFF[23]的正向数据(图8 (c))。(c)是包围来自坦克和太阳穴的数据[19]数据集(图8 (d))。(d)是镇定的室外大规模数据(图9 (a))。

对体积空间进行更有效的采样。它优化了两个网络:一个粗一个细。粗网络使用分层抽样，细网络根据粗网络的输出使用更有信息的抽样。这个过程将更多的样本分配到我们期望包含可见内容的区域。在训练阶段为所有采样点计算空间颜色先验，而为了提高效率，只计算Eq. 2)中权重(w_i大于10−3的点。

IV. 实验

为了与之前的方法进行公平的比较，我们在各种数据集上评估我们的方法:来自LLFF的正面数据[23]，来自NeRF的合成数据[24]，来自休闲3D数据集的室内周围数据[11]，自收集的室外大规模数据(表1中的“礼堂”和“剧院”)，以及来自坦克和寺庙的包围数据[19]数据集。图8显示了不同种类的不同射击轨迹

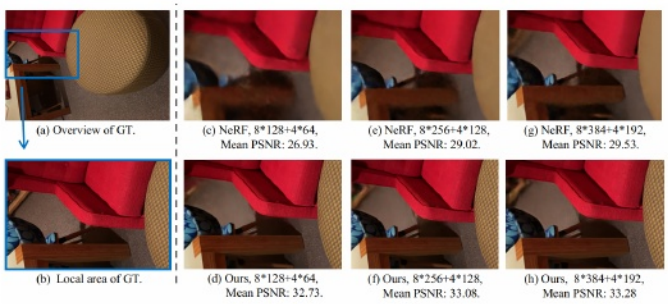

图9所示。NeRF和我们的方法在随机3D数据集“Library”中的比较[11]。两种方法使用相同的代码库，包含8 + 4个全连接ReLU层(在图5中介绍)。我们用不同的隐藏通道测试它们的性能。(a,b)是groundtruth的概述和局部区域。(c,e,g)为NeRF结果，(d,f,h)为我们的结果。PSNR值是验证集的平均结果。

的数据。接下来，通过定量和定性的评估来验证所提方法的性能。

A. 定量评估

定量评价采用PSNR、SSIM和LPIPS进行评估[35]。PSNR和SSIM值越小，精度越高，LPIPS值越高，视觉质量越好。我们将我们的方法与包括SRN[31]、NV[22]、LLFF[23]和NeRF[24]在内的先前的技术水平进行了比较，如表1所示。

对于视野范围较小的简单场景，如LLFF数据集中的“房间”和“堡垒”，NeRF在足够的内存容量下取得了良好的性能。空间颜色先验有助于揭示高频细节，改进相对较小。对于具有大规模周围视图的复杂场景，例如来自休闲3D数据集的“Library”和“Attic”[11]，由于网络大小的限制，NeRF表现不佳。所提出的方法取得了更好的性能，因为所提出的空间颜色先验有助于降低大规模场景的网络容量要求。如图9所示，随着网络规模的增长，NeRF的性能会越来越好，这说明NeRF的渲染质量受到其网络容量的限制。然而，更大的内存大小需要更多的复杂度，这就限制了

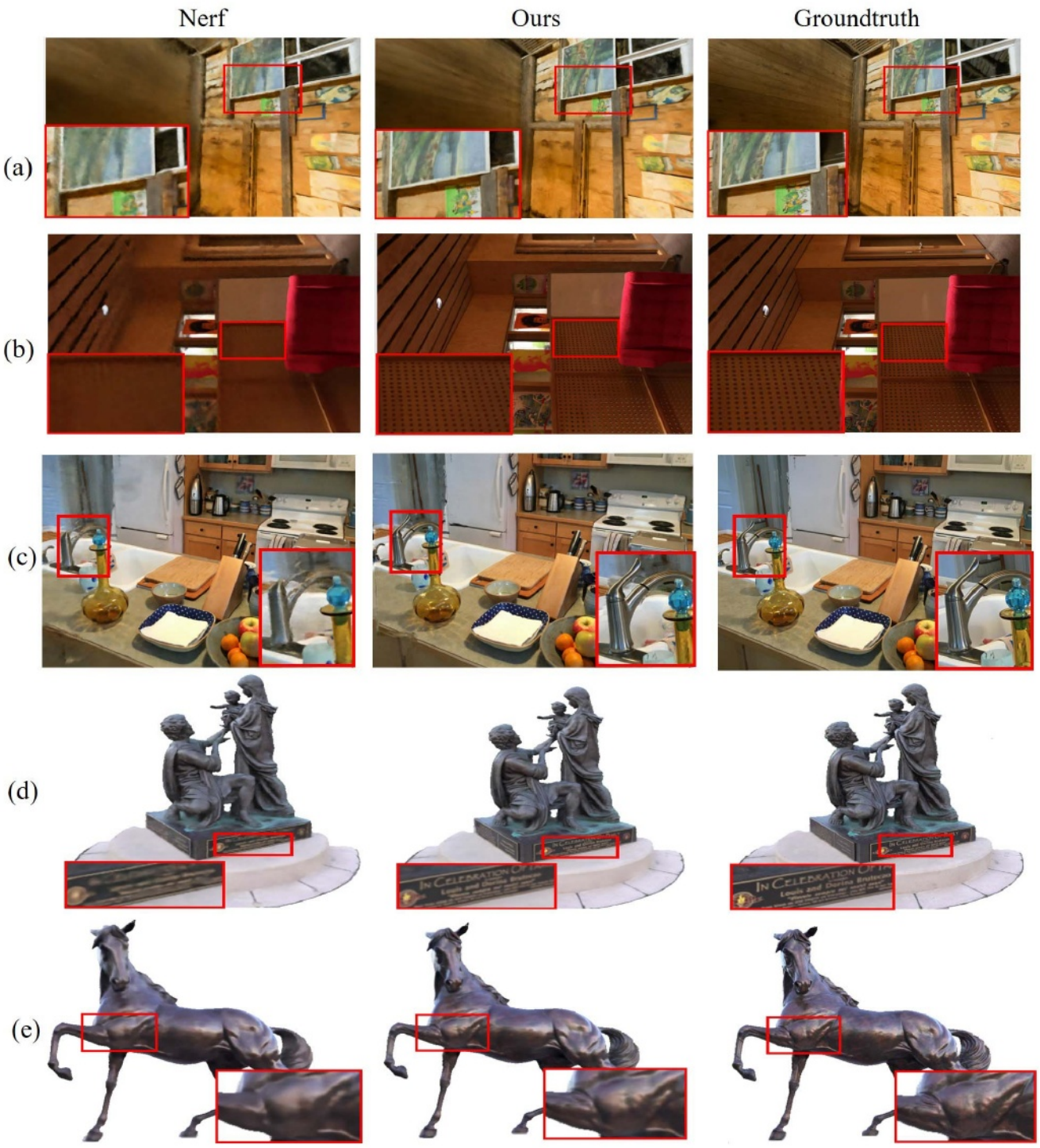

图10所示。与之前的NeRF方法[24]相比，在公共数据集上的定性评价:(a,b,c)来自随机3D [11]， (d,e)来自坦克和太阳穴[19]。实验表明，基于空间颜色先验的残差学习方案与现有方法相比，可以产生更清晰的细节。

尺寸从增加太多。网络增长带来的改善也很小。另一方面，在提出的空间颜色先验的帮助下，对网络容量的要求大大降低，我们提出的残差学习方案即使在较小的网络上也实现了更好的质量。

我们还比较了NeRF和我们提出的在不同分辨率下呈现新视图的方法的性能，如表2所示。对于更高的分辨率，我们的方法和NeRF之间的差距更大，证明了我们提出的方法在高分辨率下生成逼真渲染结果的能力。

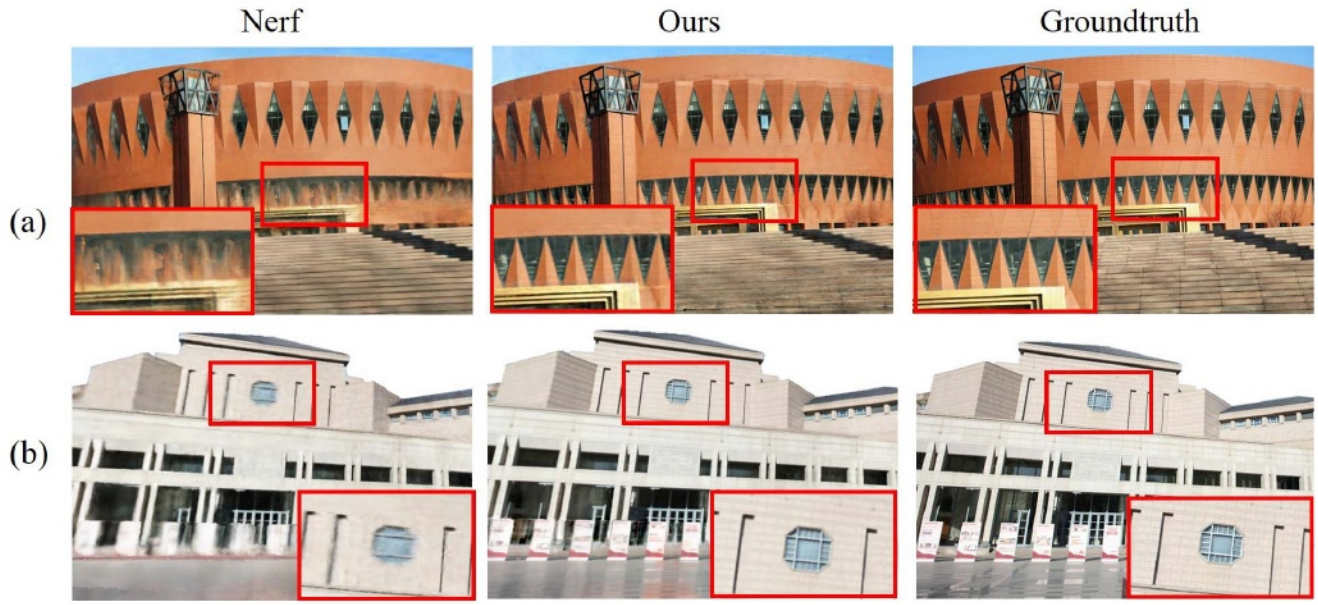

图11所示。在自收集的大型户外场景“剧院”(a)和“礼堂”(b)上，与之前的方法NeRF[24]进行定性评价。实验表明，我们的方法在大型场景中取得了比NeRF更具竞争力的性能。

表二世
nerf和我们的方法的PSNR比较不同
决议。“l library”和“a tic”是来自casual 3d[11]数据集的室内数据

	Library 960×720	Library 1200×900	Attic 592×880	Attic 886×1330
NeRF	29.02	28.09	23.44	23.64
Ours	33.08	33.21	24.93	25.25

B. 定性评估

参考颜色是根据空间颜色先验计算的。它在大部分区域接近真实的渲染结果，可能会因为像素的错误投影而在角落出现失真。残差颜色预测具有部分纠正这些问题的潜力。此外，它还可以添加不同视角的不同光线阴影(从图7中我们可以看到)。以下定性评估表明，我们提出的方法实现了鲁棒的参考颜色计算和高质量的渲染性能。

· **整体性能**。我们的方法提出了一个基于残差的框架来利用空间颜色先验，并将这一思想应用于NeRF。图10和图11为不同场景下与NeRF[24]的定性对比。对于NeRF，纹理的高频信息很难学习。它会丢失细节信息。我们的方法把高频学习任务变成了低频学习任务。残差颜色只需要记住低频信息，因为计算的参考颜色已经捕获了场景的高频纹理。因此，高频信息得到了更好的保留，我们的方法有了更清晰的细节。特别是，NeRF只能恢复有限的场景，否则质量很低。对于大规模场景，由于网络规模的限制，NeRF往往表现不佳。而我们的方法可以有效地处理大规模场景，因为提出的空间颜色先验有助于降低网络容量需求。例如，对于图10 (b,c)的复杂室内场景和图11 (a,b)的大规模室外场景，我们的结果显示了在高质量渲染方面的显著提升。图12显示了与SRN[31]、NV[22]、NeRF[24]和新发表的NSVF[21]的定性比较。实验表明，我们的方法取得了比以前的方法(SRN)更好的性能。

NV和NeRF)和与NSVF相当的性能。 · **遮挡处理**。参考颜色由投影像素计算。如果有遮挡，错误的投影像素可能会影响参考图像的质量。我们应用特征过滤器和联合训练来处理这种限制。图13 (b)显示，在没有遮挡检测的情况下，由于像素的错误投影，参考颜色会出现明显的伪影。由于像素的错误投影，参考颜色会出现明显的失真。残差颜色预测具有部分纠正这些问题的潜力，但补救措施不能是完美的，所产生的图像仍然受到某些伪影的影响。通过我们的patch特征过滤器和联合训练，参考颜色的计算不受遮挡的影响，如图13 (c)所示。同时，如图14所示，性能增益主要来源于残差学习方案。特征过滤器的作用是处理遮挡。因此，它需要与联合训练相结合。

C. 局限性

尽管所提出的方法可以生成高质量的新视图，并优于现有的最先进技术

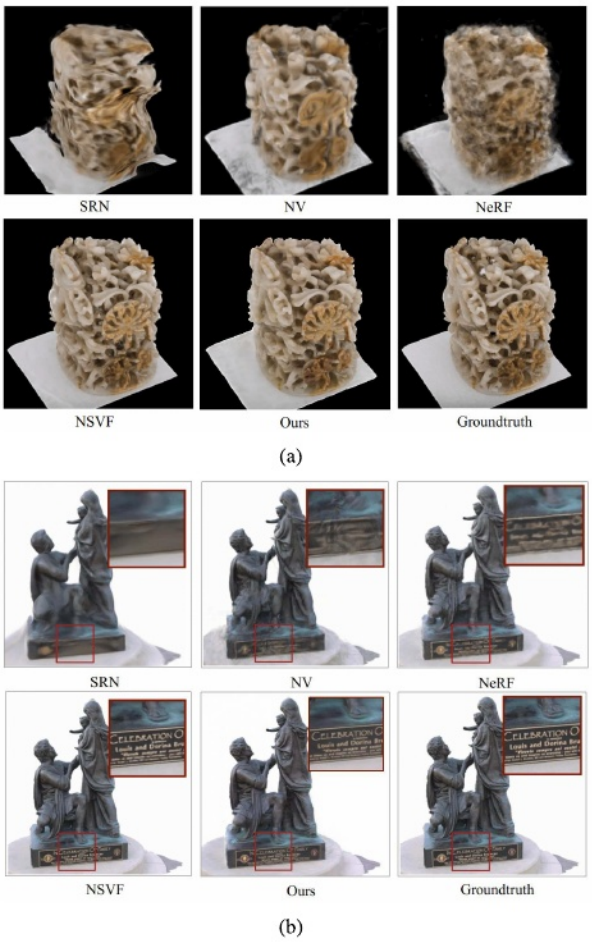

图12所示。对BlendedMVS数据集[34]中的“Jade”(a)和坦克和寺庙数据集[19]中的“Family”(b)进行定性评价，并与SRN[31]、NV[22]、NeRF[24]和新发表的NSVF[21]方法进行比较。实验表明，我们的方法取得了比以前的方法(SRN、NV和NeRF)更好的性能，与NSVF性能相当。

方法,反映地区依然贫穷的性能。残差颜色学习方案将朗伯曲面上的高频学习任务转化为低频学习任务。然而，在反射区域，不同视角的观测值差异很大，因此所提出的基于残差的方法并不比NeRF具有优势。从图15中我们可以看到，我们提出的方法对朗伯表面效果最好，而对反射表面的NeRF效果相似。在高光区域保持精确的细节仍然是一个挑战。

拟议的空间剩余学习颜色先验是有效的和有效的。它对于神经渲染是鼓舞人心的，可以很容易地引入到其他框架中。同时,提出额外的训练失去光辉的颜色增加了计算复杂度主要用于去除小工件由阻塞引起的。额外的计算负担是低效的性能改善。没有额外的亮度颜色MLP，我们提出的残差学习方案在大多数领域仍然优于NeRF，而只是在遮挡中引入小的伪影。神经呈现的现有方法仍然受到渲染质量差,实际使用中必须解决的,所以这一点

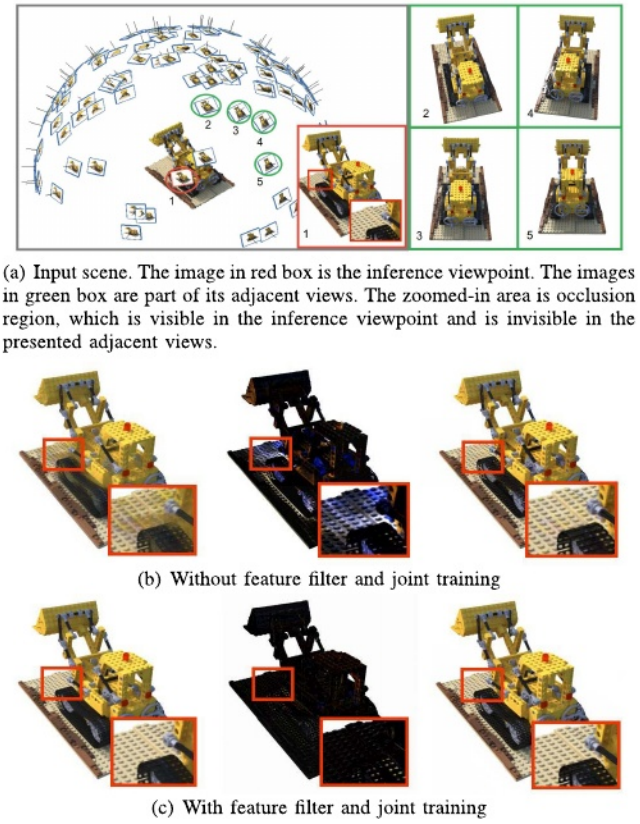

图13所示。分别使用或不使用特征过滤器和联合训练渲染图像的对比。(b,c)从左到右:参考图像、残差图像和结果图像。(b)的参考彩色图像有明显的工件由阻挡投影像素,而参考彩色图像(c)是更准确。因此，通过特征滤波和联合训练，残差彩色图像不需要弥补错误，结果图像在遮挡方面有改善。

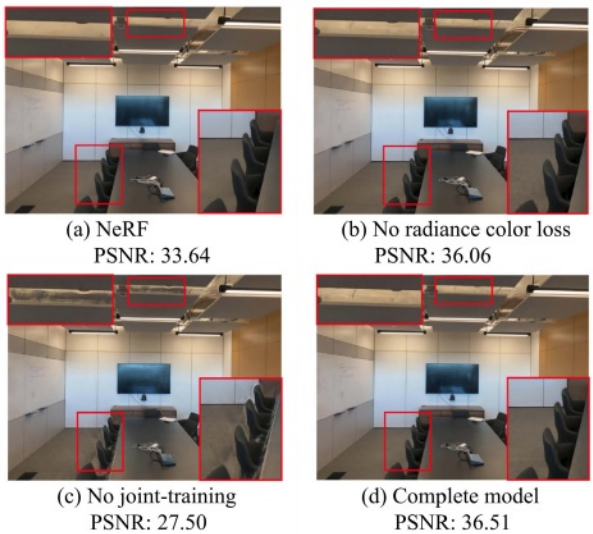

图14所示。(a)是原始NeRF的结果。(b)使用残差学习来利用空间颜色先验，而没有额外的亮度颜色损失。它已经达到了比NeRF更好的性能，而红色框表示遮挡区域模糊(顶部框)或有重影。(c,d)在残差学习方案之外引入额外的亮度颜色损失。(c)使用特征过滤器而不进行联合训练，这意味着辐亮度颜色和残差颜色使用自己的密度预测。(d)同时使用特征过滤器和联合训练。对比表明，性能提升主要来自残差学习方案。而特征过滤器必须与联合训练相结合，以去除遮挡引起的伪影。

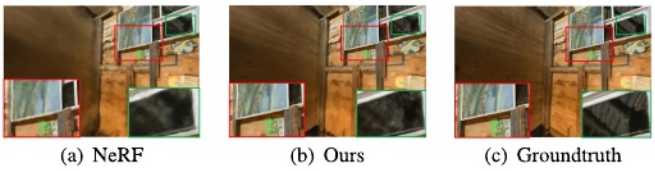

图15所示。朗伯面(红块)与反射面(绿块)的对比。该图显示，我们基于残差的方法在朗伯表面上优于NeRF，而对于反射表面，我们的方法与NeRF相似。

论文主要专注于合成高质量的新视图，而不关心复杂性。未来，我们将在我们的残差学习框架中探索更有效的方法来处理遮挡。

V. 结论

本文旨在改善自由移动摄像机合成新视图的沉浸式体验，认为试图使用全连接网络记忆环境的纹理细节和几何形状的传统场景表示网络在实践中无法保留高频细节，并提出了一个新的框架，通过使用所提出的空间颜色先验作为辐射度颜色预测的参考来学习残差颜色。实验表明，所提出的方法取得了比以前的技术更令人愉快的视觉结果，特别是对于包含复杂纹理和大表面积的环境。所提出的方法对朗伯曲面效果最好，仅对非兰伯量曲面达到与先前方法相当的性能。检测等领域使用分割方法[3]可能有助于这一挑战,还将为未来的调查。

参考文献

[1] K.-A. Aliev, A. Sevastopolsky, M. Kolos, D. Ulyanov, and V. Lempitsky, "Neural point-based graphics," in *Proc. 16th Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, Aug. 2020, pp. 696–712.

[2] R. Anderson *et al.*, "Jump: Virtual reality video," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–13, Nov. 2016.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convo-lutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[4] P. Debevec, Y. Yu, and G. Borshukov, "Efficient view-dependent image-based rendering with projective texture-mapping," in *Proc. Eurograph. Workshop Rendering Techn.* Springer, 1998, pp. 105–116.

[5] A. Fitzgibbon, Y. Wexler, and A. Zisserman, "Image-based render-ing using image-based priors," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 141–151, 2005.

[6] J. Flynn *et al.*, "DeepView: View synthesis with learned gradient descent," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2367–2376.

[7] Y. Furukawa and C. Hernández, "Multi-view stereo: A tutorial," *Found. Trends Comput. Graph. Vis.*, vol. 9, nos. 1–2, pp. 1–148, 2013.

[8] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," 2020, *arXiv:2002.10099*.

[9] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[10] P. Hedman, S. Alsisan, R. Szeliski, and J. Kopf, "Casual 3D photogra-phy," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–15, Nov. 2017.

[11] P. Hedman, S. Alsisan, R. Szeliski, and J. Kopf, "Casual 3D photogra-phy," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 234:1–234:15, Nov. 2017.

[12] P. Hedman, T. Ritschel, G. Drettakis, and G. Brostow, "Scalable inside-out image-based rendering," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–11, Nov. 2016.

[13] J. Huang *et al.*, "Adversarial texture optimization from RGB-D scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1559–1568.

[14] M. Irani, T. Hassner, and P. Anandan, "What does the scene look like from a scene point?" in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2002, pp. 883–897.

[15] J. T. Kajiya and B. P. Von Herzen, "Ray tracing volume densities," *ACM SIGGRAPH Comput. Graph.*, vol. 18, no. 3, pp. 165–174, Jul. 1984.

[16] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 365–376.

[17] O. Kähler, V. A. Prisacariu, C. Y. Ren, X. Sun, P. Torr, and D. Murray, "Very high frame rate volumetric integration of depth images on mobile devices," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 11, pp. 1241–1250, Nov. 2015.

[18] M. Klingensmith, I. Dryanovski, S. S. Srinivasa, and J. Xiao, "Chisel: Real time large scale 3D reconstruction onboard a mobile device using spatially hashed signed distance fields," in *Robotics: Science and Systems*, vol. 4, no. 1. Citeseer, 2015.

[19] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, Jul. 2017.

[20] C. Lassner, "Fast differentiable raycasting for neural rendering using sphere-based representations," 2020, *arXiv:2004.07484*.

[21] L. Liu, J. Gu, K. Z. Lin, T.-S. Chua, and C. Theobalt, "Neural sparse voxel fields," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 15651–15663.

[22] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh, "Neural volumes: Learning dynamic renderable volumes from images," 2019, *arXiv:1906.07751*.

[23] B. Mildenhall *et al.*, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–14, Aug. 2019.

[24] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 405–421.

[25] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3504–3515.

[26] K. Park *et al.*, "Nerfies: Deformable neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 5865–5874.

[27] E. Penner and L. Zhang, "Soft 3D reconstruction for view synthesis," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–11, Nov. 2017.

[28] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5301–5310.

[29] C. Reiser, S. Peng, Y. Liao, and A. Geiger, "KiloNeRF: Speed-ing up neural radiance fields with thousands of tiny MLPs," 2021, *arXiv:2103.13744*.

[30] G. Riegler and V. Koltun, "Free view synthesis," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 623–640.

[31] V. Sitzmann, M. Zollhöfer, and G. Wetzstein, "Scene representation networks: Continuous 3D-structure-aware neural scene representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1121–1132.

[32] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik, "Multi-view supervision for single-view reconstruction via differentiable ray consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2626–2634.

[33] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "ElasticFusion: Real-time dense SLAM and light source estimation," *Int. J. Robot. Res.*, vol. 35, no. 14, pp. 1697–1716, Sep. 2016.

[34] Y. Yao *et al.*, "BlendedMVS: A large-scale dataset for generalized multi-view stereo networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1790–1799.

[35] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unrea-sonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[36] D. Zhong, L. Han, and L. Fang, "IDFusion: Globally consistent dense 3D reconstruction from RGB-D and inertial measurements," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 962–970.

[37] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," 2018, *arXiv:1805.09817*.

韩磊毕业于中国香港科技大学和清华大学，主修电气工程。2013年7月获清华大学学士学位，2020年获香港科技大学博士学位。目前是华为海思硅的工程师。他的研究重点是多视点几何和3 d计算机视觉。

钟大伟在清华大学清华-伯克利深圳研究院(TBSI)学习数据科学与信息技术。2019年7月获得同济大学学士学位。他目前的研究是关于3 d计算机视觉。

林丽，2010年获浙江大学学士学位，2015年获浙江大学博士学位。她目前是华为海思的工程师。

郑凯，2016年毕业于中国威海哈尔滨工业大学，获学士学位;2018年毕业于中国哈尔滨哈尔滨工业大学，获计算机专业硕士学位。他目前是海思的研发工程师。他的研究兴趣包括智能产业，3 d重建,汽车车辆。

吕方，IEEE高级会员，2007年获中国科技大学理学学士学位，2011年获香港科技大学博士学位。现任清华大学电子工程系副教授。她的研究兴趣包括计算成像和视觉智能。她目前是IEEE图像处理汇刊和IEEE多媒体汇刊的副主编。