用于三维场景的广义语义分割的神经语义域

苏哈尼沃拉[1] *诺哈拉德万[1] *克劳斯Greff[1]亨宁迈耶[1]Kyle热那亚[1]
Mehdi S。M.萨贾迪[1]艾蒂安锅[1]安德里亚塔利亚萨奇[1,2]丹尼尔达克沃斯[1]

[1]谷歌研究[2]多伦多大学

摘要

我们提出了NeSF，一种单独从提出的RGB图像产生3D语义域的方法。取代经典的三维表示，我们的方法建立在最近的隐式神经场景表示的工作上，其中三维结构被点智慧函数捕获。我们利用这种方法来恢复三维密度场，然后在此基础上训练一个由假设的二维语义映射监督的三维语义分割模型。尽管仅在二维信号上进行训练，但我们的方法能够从新的摄像机姿态中生成3D一致的语义映射，并可以在任意的三维点上进行查询。值得注意的是，NeSF与任何产生密度场的方法都可以兼容，其精度随着密度场质量的提高而提高。我们的实证分析表明，在复杂的、真实渲染的合成场景上，其质量与竞争性的二维和三维语义分割基线相当。我们的方法是第一个提供真正密集的三维场景分割，只需要2D监督来进行训练，并且不需要任何语义输入来对新场景进行推理。我们鼓励读者访问这个项目的网站。

1.介绍

对数字图像和视频所捕获的三维场景的高级语义理解是计算机视觉的一个基本目标。研究了场景分类[72]、对象检测[81]、语义分割[75]、实例分割[49]等任务，从RGB和其他传感器中推断出场景的语义描述，为视觉导航[7]和机器人交互[6]等应用奠定了基础。

场景理解最常见的方法是将范围缩小到二维（图像空间）推理，其中经典的图像到图像架构[118]被训练

*表示相等的贡献。

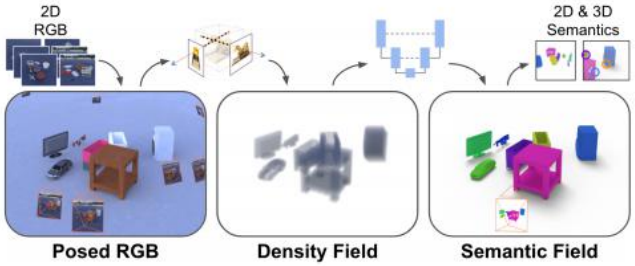

图1。概述-我们在提出的2D RGB图像和2D语义映射的集合上训练我们的方法，每个集合描述一个独立的场景。给定一组新的二阶RGB图像，我们提取了场景三维几何的隐式体积表示，并推断出一个三维语义场。语义字段可以用来从新的相机姿态中渲染密集的二维语义映射，或者直接在三维中查询。我们的方法可以推广到新的场景，并且在训练时每个场景只需要一个语义映射。我们鼓励读者访问该项目的网站。

大量带有语义注释的图像集合[78]。这些方法虽然应用起来很简单，但它们只生成每个像素的注释，并且在很大程度上忽略了场景的底层3D结构。相反，我们的目标是使用一组具有已知姿态的RGB图像来产生一个三维语义域：一个将3D位置映射到语义类别上的概率分布的函数。

对于三维语义分割，以往的工作依赖于三维传感器[36,59,71]和/或三维语义标签[9,30]。针对三维点云[23,112]、体素网格[131]和多边形网格[51,87]设计了卷积网络体系结构。然而，3D传感器并不像它们的RGB相机那样便宜，也不那么广泛使用，而且3D注释比2D传感器更具挑战性，因此通常缺乏[41]。

为了克服这一挑战，研究人员采用了混合2D/3D推理，将密集监督的语义信号传播回潜在的3D底物[31,41,73]。在测试时，这些方法仍然是如此

需要提供一个经典的三维表示作为输入，因此限制了它们的适用性和性能。一个有趣的例外是Atlas [97]，它只需要在测试时摆姿势的照片，但仍然需要3D监督来训练。

与这些发展并行的是，一个基于隐式的、协调的函数的新的三维表示家族也出现了[19,91,105]。在这种情况下，训练神经网络预测数量，如占用、符号距离、密度和辐射。与几何的显式表示不同，神经隐式方法具有记忆效率，并且能够捕获令人印象深刻的几何细节水平[128]。然而，建立在这种方法上的绝大多数方法都是针对计算机图形应用程序，如新的视图合成[80,114,136,142,146]，而没有对场景的可解释的语义理解。一个显著的例外是语义-nerf[149]，它除了辐射和密度外，还回归了每3d点的语义类。与NeRF [94]类似，该方法仅适用于同一场景中的新视图，并不提供经典语义分割中所期望的泛化形式：在新场景上推断语义的能力。

我们介绍了神经语义域（NeSF），这是一种通过图像空间语义注释对三维场景进行语义分割的方法；见图1。与语义-nerf不同，NeSF概括到在训练时未观察到的场景。NeSF代替显式场景表示，建立在NeRF等方法恢复的几何隐式表示之上。特别地，我们将一个神经网络应用于NeRF的密度场来得到我们所称的场景的语义场。因此，语义域被定义为一个坐标函数映射到语义类别上的概率分布。与语义-nerf类似，我们应用体积渲染方程来生成二维语义映射，允许在图像空间中对已提出的语义注释进行监督。据我们所知，NeSF是第一个能够仅从提出的RGB图像中产生密集的二维和三维新场景分割的方法。仅从2D监督中推断3D信息的能力对于大规模部署3D计算机视觉至关重要；虽然2D传感器无处不在，但3D传感器价格昂贵、笨重，而且不太可能在大众市场上部署。

由于三维语义标注场景的大规模数据集稀缺，我们提出了三个日益复杂的新数据集： KLEVR、ToyBox5和ToyBox13。虽然用于二维和三维语义场景理解的数据集已经存在[10、12、30、125]，但它们缺乏同时评估二维和三维语义分割所需的规模、细节、真实性和精度。我们构建了超过1000个随机放置的玩具大小的物体场景，并使用真实的灯光和材料渲染数百张的RGB图像。值得注意的是，随机对象的放置破坏了可用数据集中存在的对象之间的关系一致性，

启用更困难的任务。每个RGB图像都与相应的地面真实相机的内部图像和外部图像、语义图和深度图配对。我们在这三个数据集上评估我们的方法，并比较其性能，以强迫

–

在2D和3D场景理解中的直观技术。贡献。

我们介绍了第一种方法，为仅在姿态RGB图像和语义映射上训练的新场景生成三维语义域。与之前的工作不同，我们的方法(i)可以在一个有界的三维体积内的任何地方查询，（ii）能够从新的摄像机姿态呈现语义映射，以及（iii）推广到新的场景，每个场景的训练时间只有一个语义映射。

我们提出了三个新的合成数据集的二维和三维语义场景理解。总的来说，这些数据集包括超过1000个场景和300万个真实渲染和语义注释的帧框架。在发布后，这些数据集将与代码一起发布到社区，以复制它们。

2.相关工作

我们现在简要概述了语义分割[52,96]和三维重建[64]的相关工作。

语义分割语义分割是一个被广泛研究的领域，大多数方法都是针对一个完全监督的单一模态问题（2D [4,16,17,22,83,119]或3D [14,21,29,47,82,95,147,150]）。像DeepLab [16]这样的二维方法训练一个CNN来分割图像中的每个像素。对于各种形状表示，也有类似的三维方法——点云[56,110,112,113,122,129,135]，稀疏或密集的体素网格[24,25,31,43,50,117,123]，或网格[51,57]。与这些方法相比，我们的方法仅从2D输入和监督中重建并分割了一个密集的3D表示，并且不需要地面真实的3D注释或输入几何图形。

混合和多模态方法。许多方法使用一种数据模式来监督或通知另一种数据模式[1、3、37、38、42、48、65、68、69、76、93、98、104、130、148]。对于三维语义分割，多视图融合[2,55,73,84,86,89,133,133,145]是一种流行的只需要图像监督的方法。然而，这些方法只在图像领域进行推理，需要一个输入的三维基底，如点云或多边形网格来聚合二维信息。同样，Genova等人。[41]提出了一种从二维监督中分割三维点云的方法，但仍然需要输入三维几何图形。在另一项工作中，研究人员提出了受益于2D图像特征[31,62,73,74,134]。与我们的方法不同，这些方法还需要一个完整的3D监督。

隐式表示。与我们的方法最相似的是，Atlas [97]学习了一个三维隐式TSDF重建

二维图像，同时也学习分割预测的场景几何形状。然而，这种方法需要地面真实的三维数据和监督，而我们的方法只需要在训练和测试时间的图像。其他方法使用隐式表示来重建三维场景[109,126]或形状[11、18、20、33、34、40、92、101、103、106、108、111、120、128]。最近的一项研究只用图像监督来处理这个问题[126]，但没有考虑语义。

神经辐射场。近年来，多种基于NeRF [94]的方法在新视图合成
[5,39,53,63,77,79,80,85,99,107,115,124,136,141]、3D重建
[8,8,15,27,35,54,60,61,102,116,132,139,142,143]、生成建模[70,90,100,121]和语义分割方面得到流行[149]。这些模型中的大多数在新的视图合成上显示了令人印象深刻的结果，但只适用于单场景设置。其他人则可以概括到新的场景，但却专注于新的视角的合成或重建。相比之下，我们的方法是第一个能够在没有监督的情况下生成新场景的三维语义分割的方法。

3.方法

我们在S场景集合上训练NeSF，每个场景都由RGB图像集合{C描述$^{gt}_{s}$,c∈ [0, 1]$^{HW\ 3\times\times}$并与语义映射集合{S∈$Z^{gt}_{s,c+}{}^{H\times W}$}.
图像和地图都由相机索引c和场景索引s进行索引。为了便于说明，我们假设每个RGB映射都与一个语义映射配对，并且每个场景都包含C个映射对，但该方法本身并没有这样的假设。与之前的工作类似，我们也假设了相机校准参数{V的可用性s,c$^{\Gamma}$∈R}提供3D场景中每个像素和投射的3D射线之间的显式连接。我们考虑了联合估计摄像机校准和场景表示的问题，见[77,137,140]。我们的方法包括两个阶段
将在以下小节中进行了描述：

第3.1节：在第一阶段，我们对提出的RGB地图进行预训练{(C^{gt}_{s},c，Vs,c*独立为每个场景的∈[1・・・S]*。这就产生了一组具有网络参数{的神经辐射场θ_s}.为了专注于理解三维几何的核心任务，我们忽略了这些场的辐射部分，而使用体积密度场(x |$\sigma\theta_s$)∈[0，∞)就在下面。

第3.2节：在第二阶段，我们训练一个由τ = {τ参数化的密度-语义翻译网络Tunet， τmlp}.给定一个由密度场表示的场景的三维几何图形$\sigma_s\sigma$= （・ |θ_s)，该网络产生一个三维语义字段(x |σ_s，τ)分配每个点在语义类别上的概率分布。当平移网络产生一个三维场时，我们应用体积渲染方程得到二维语义-
抽搐地图从参考相机姿态{Vs,c}.然后，可以将预测的语义映射与它们的地面真实映射进行比较$^{gt}_{s}$,c}以一种可区分的方式出现。

3.1.NeRF预训练

为了提取每个场景的三维几何图形的精确、密集的表示，我们利用了[94]中提出的神经辐射场。为了简化表示法，我们删除了场景索引
因为所有场景都可以独立和并行训练。更具体地说，给定一组姿势的RGB图像{(C，V^{gt}_{c}c，用r R (V表示~$c^{gt}_{c}\theta$)，通过最小化平方光度重建损失，训练一个带有参数的神经辐射场模型：

$$L_{rgb}(e) = 对 E_{\mathbf{rR}(\sim c)}[||C(r\ |\ )\theta^{gt}_{c} — C(r)||^{2}_{2}\rfloor (1)$$
c

$^{gt}_{c}\theta$其中C (r)是通过图像C中一个像素的光线的地面真实颜色，颜色C（r |）是通过应用射线的远近边界∈的体积渲染方程来计算的n, tf]:

$$\mathcal{C}(\mathbf{r}\,|\,\theta) = \int_{t_n}^{t_f} w(t\,|\,\theta)\cdot \mathbf{c}(t\,|\,\theta)\,dt \qquad (2)$$

设r (t)=o + td表示沿原点o和方向d的射线的点。权重w (t)=w（r(t)）随后定义为：

$$w(t) = \underbrace{\exp\left(-\int_{t_n}^{t}\boldsymbol{\sigma}(u)\,du\right)}_{\text{visibility of }\mathbf{r}(t)\text{ from }\mathbf{o}}\cdot\underbrace{\boldsymbol{\sigma}(t)}_{\text{density at }\mathbf{r}(t)} \qquad (3)$$

其中，体积密度(t)和辐射场c (t)是由一个多层感知器（i。σe.MLP)与傅里叶特征编码。我们向读者参考原始的工作[94]，以获得进一步的细节和这些积分[88]的离散化。

训练虽然神经辐射场被认为训练速度很慢，但我们发现，我们能够在谷歌云平台上的8个TPUv3核上，在20分钟内将单个模型匹配到足够的质量。≈一旦训练好了，每个场景的参数就会变成{θ_s}是固定的。

.2.3语义推理

我们现在提出了一种将三维密度场映射到三维语义域的方法。$\sigma\sigma\sigma$综上所述，我们训练一个翻译模型T（| τ）来产生一个语义域（x |，τ），它被允许访问场景的密度场，其中s在空间的每个三维点为语义类别分配一个概率分布。我们仅使用二维注释来优化翻译模型的参数τ。我们的灵感来自于将几何的显式表示转换为语义[32,112]的方法，并在观察它的目的上

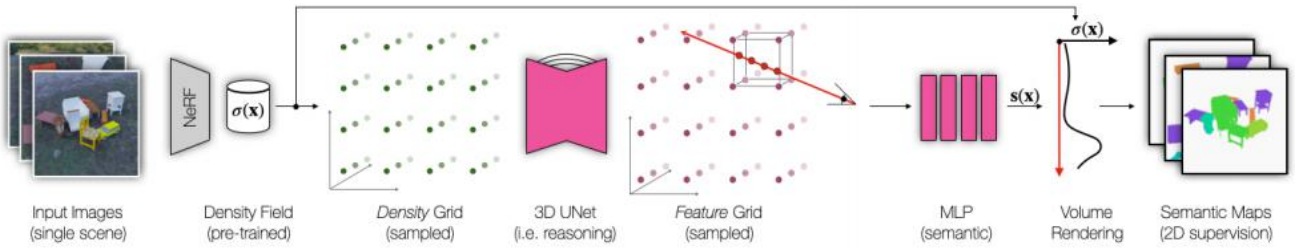

图2。架构-给定一个预先训练过的NeRF模型，我们对其体积密度网格进行采样，以获得三维场景表示。通过使用完全卷积的卷积网络，该网格被转换为语义特征网格，从而允许几何推理。语义-特征网格利用体积呈现方程将其转化为语义概率分布。注意，语义3D UNet是在训练场景集中的所有场景中进行训练的，尽管为了简单起见并没有明确地描述。此外，请注意，NeSF仅使用二维监督信号进行训练，并且在测试时没有提供分割图。

密度场提供了一个隐含的几何概念。为了便于说明，图2所示的整体架构被分解为几个离散的步骤： 01密度网格提取、02空间推理、03特征解码、04监督和05数据增强。

密度网格提取。我们的方法首先要均匀地评估样品间间距为e的三维晶格上的密度场：

$$\Sigma_S = (\Sigma\sigma_S) = \{(x \mid \sigma\theta_S) \text{ s.t. } x \in [-1 : e : +1]^3\} \quad (4)$$

虽然这种操作限制了原始密度场的空间分辨率，但它为进一步处理提供了一种自然的表示。

空间推理（3D）。我们应用一个3D UNet [26]到Σ_S*来获得一个特征网格*F_S具有相同的空间分辨率Σ_S:

$$F_S = F(\Sigma_S \mid \tau_{unet}) = UNet3D(\Sigma_S \mid \tau_{unet}) \quad (5)$$

σ这一步是必不可少的，因为密度场(x)的点级测量不包含足够的信息来捕获三维结构。σ毕竟，(x)只测量了某个点上的体积密度，而三维结构则需要对局部空间邻域进行推理。请注意，我们共享翻译网络参数τ_{unet}跨场景，使泛化到在训练时不可用的新场景。

功能解码。给定一个查询点$x \in R^3$，我们在特征网格F内进行插值s得到一个对应于x的特征向量。然后，我们使用一个神经网络解码器D来生成一个在语义类别上的概率分布领域：

$$s(x \mid F_S, \tau_{mlp}) = D(TriLerp(x, F_S) \mid \tau_{mlp}) \quad (6)$$

其中D是一个具有可训练参数τ的多层感知器mlp和TriLerp采用类似于[80]的三线性插值。与它们的UNet对应物一样，参数τ_{mlp}*在所有场景中共享*。

监督。为了监督参数τ的训练，我们使用了NeRF [94]中的体积渲染，但将其调整于语义-NeRF[149]中的语义映射：

$$\mathcal{S}(\mathbf{r} \mid \boldsymbol{\sigma}_s, \boldsymbol{\tau}) = \int_{t_n}^{t_f} w(t \mid \boldsymbol{\sigma}_s) \cdot \mathbf{s}(\mathbf{x} \mid \boldsymbol{\sigma}_s, \boldsymbol{\tau})\, dt \quad (7)$$

我们通过最小化呈现的语义和地面真实语义映射之间的softmax交叉熵以及一个平滑正则化术语来监督τ的训练过程：

$$\mathcal{L}_{\text{sem}}(\boldsymbol{\tau}) = \mathbb{E}_s\left[\mathcal{L}_{\text{sem}}(\theta_s, \boldsymbol{\tau})\right] \quad \text{where:} \quad (8)$$
$$\mathcal{L}_{\text{sem}}(\theta_s, \boldsymbol{\tau}) = \sum_c \mathbb{E}_{\mathbf{r}\sim\mathcal{R}(\gamma_c)}\left[\text{CE}(\mathcal{S}(\mathbf{r} \mid \theta_s, \boldsymbol{\tau}) - \mathcal{S}_c^{\text{gt}}(\mathbf{r}))\right]$$

我们包括了一个额外的光滑正则化术语，以鼓励在当地社区进行类似的预测。~我们对点x均匀性进行采样^3)和正态分布的噪声e N（0、0.01），~

$$L_{reg}(\tau) = E_{x,e}[||s(x|F, \tau) - s(x + e|F, \tau)||_2^2] \quad (9)$$

因此，我们的总损失是$L(\tau) = L_{sem}(\tau) + \lambda L_{reg}(\tau)$.

数据增强为了增加我们的方法的鲁棒性，与经典方法类似[112,113]，我们以围绕z轴的随机旋转的形式应用数据增强(i。e.向上地特别是，我们在训练的每一步中随机抽取一个角度V 2 [0,2)。π我们没有在NeRF的原始坐标系中提取一个密度网格，而是构造了一个旋转变换R，并在点x处查询NeRF$^{\backslash}$= R (x)，导致以下结果密度网格：

$$\tilde{\Sigma}_S = \{(x\sigma^{\backslash} \mid \theta_S) \text{ s.t. } x \in R^{-1}([-1 : e : +1]^3))\} \quad (10)$$

请注意，这个过程并不需要对NeRF模型进行再训练。

	克莱夫尔	玩具盒5	玩具盒13
#场景	100 / 20	500 / 25	500 / 25
# cameras/scene	210 / 90	210 / 90	210 / 90
#全摄像头	36,000	1,575,00	1,575,00
帧分辨率	256 256×	256 256×	256 256×
# objects/scene	412	412	412
#对象实例	5	25,905	39,695
#背景实例	1	383	383

表1。数据集统计数据-每个数据集由一组火车组成
以及新奇的场景，其中每个场景的摄像机被分割成一个火车和测试集（用一个“/”表示）。

图3。数据集示例——每一帧都包括一个RGB图像、语义图和深度图（如图所示）。

4. 数据集-表1和图3

为了研究NeSF，我们需要从多个角度来描述大量场景的外观和语义的数据集。虽然现有的基于室内和自动驾驶传感器捕获的数据集已经存在[10,12,30,127]，但我们希望有一个受控的设置，这样可以消除运动模糊、相机校准错误和物体运动等干扰物。为此，我们引入了构建在Kubric [45]上的三个新的数据集： KLEVR、ToyBox5和ToyBox13。每个数据集由数百个合成场景组成，每个场景都包含随机放置的3D对象，这些对象由路径跟踪器[28]逼真地渲染，支持软阴影、物理材料和全局照明效果。每个场景由一组姿势帧描述，其中每个帧提供一个RGB图像、一个语义图和一个从共享相机姿势渲染的深度图。我们提供了Kubric工作脚本来生成这样的场景，以便使后续研究能够构建更具挑战性的数据集。

克莱夫。我们设计的KLEVR数据集是一个简单的测试台，类似于机器学习中的MNIST。灵感来自
CLEVR [66]，每个场景包含4到12个简单的几何物体，随机放置在哑光灰色地板上。每个对象都被分配了一个随机的色调和比例，并被限制在一个固定的边界框内。一个对象的语义类别被设置为等于它的几何类。立方体、圆柱体等)。虽然只有每个物体的形状在语义上相关，但颜色、比例和位置作为干扰物。对于每个场景，我们从针对场景原点的随机采样的相机姿态中呈现300帧。相机的姿势被限制在场景周围的上半球。对于每一帧，我们都会渲染一个RGB图像、一个语义映射和一个深度映射。

玩具盒5和玩具盒13。这些数据集被设计用来模仿儿童卧室的场景，并被设计得更具挑战性。场景是由大量的ShapeNet [13]对象的词汇表和在野外捕获的HDRI背景（地板、地平线和环境照明）构建的[144]。和KLEVR一样，每个场景由412个随机放置的物体组成，帧也被渲染从300个独立采样的相机姿势中提取。分别从ToyBox5和ToyBox13的5个和13个最常见的对象类别中随机采样对象。这种分裂在3D深度学习文献[33,40,46,92]中被使用。就像对象本身一样，背景在构建一个场景时也是随机采样的。由于每个类别有数千个对象可供选择，因此大多数对象实例很少或只出现一次。

列车/测试分离。为了从同一场景中的新视图进行评估，我们将每个场景的帧随机划分为训练摄像机和测试摄像机；后者代表通常用于评估新视图合成[94]中的方法的集合。为了跨场景进行评估，我们将场景进一步划分为火车场景和
小说场景。

5. 实验

我们在Sec-中描述的三个数据集上评估NeSF
tion 4.除非另有说明，我们在所有火车场景的所有火车摄像机上训练NeRF模型。为了模拟标签稀缺的机制，我们选择从每个场景中随机选择的9个摄像机对应的语义地图中提供NeSF监督。对于二维评估，我们从每个新场景的训练摄像机中随机选择4个摄像机。对于三维评估，我们使用摄像机参数和地面真实深度图从相同的4个摄像机中获得一个标记的三维点云。语义分割根据二维和三维的平均交叉过并集进行评估。进一步的细节见补充材料。

培训详细信息。每个场景都通过训练一个独立的NeRF25k步，Adam的初始学习速率为1e-3，根据余弦规则衰减到5.4e-4。我们的NeRF架构遵循了最初的工作。NeSF使用Adam进行5k步的训练，初始学习速率为1e-3，衰减到4e-4。作为NeSF的输入，我们通过e=1/32密集探测来离散密度场

	训练摄像机2D mIoU	3D mIoU	测试摄像机2D mIoU	3D mIoU
NeSF	92.7	97.8	92.6	97.5
DeepLab [16]	97.1	N/A	N/A	N/A
SparseConvNet [44]	N/A	99.7	N/A	99.7

表2。KLEVR-NeSF的定量比较与二维和三维基线具有竞争力。更多细节见表3。

导致64^3[-1；+1]中的均匀间隔点^3.然后这个密度网格由,,等人的3D UNet结构处理。[26]在每个降采样阶段都有32、64和128个通道。对语义潜在向量进行了处理由一个多层感知器组成，由2个由128个单元组成的隐藏层。我们的模型是在32个TPUv3核上进行的训练。

分割基线（2D/3D）。我们将NeSF与两种流行的语义分割基线DeepLab [16]和稀疏通信vnet[44]进行了比较。DeepLab遵循传统的二维语义分割管道，从RGB图像中生成分割地图。我们在16个宽ResNet3芯片上训练DeepLab[138]v355k步。稀疏net是一种点云分割方法，与NeSF和DeepLab不同，需要明确的三维监督。我们在20 NVIDIA V100gpu上异步训练，使用基本学习率1.5e-2，在最后250k步中衰减到0。详情请参考第5.1节和补充资料。

5.1.与基线比较

我们的第一组实验评估了我们提出的方法与在KLEVR、ToyBox-5和ToyBox13数据集上的其他基准测试方法相比的性能。据我们所知，NeSF是第一种能够在推理时直接从姿态的RGB图像中同时生成三维几何、二维语义映射和三维语义标签的方法。与之前的工作[97]不同，我们的方法是在二维监督上训练的。由于目前的方法没有直接的可比性，我们将NeSF与二维图像分割和三维点云分割的竞争基线进行了比较。

与DeepLab [16]（2D）的比较。为了在二维中保持公平的比较，我们训练NeSF和DeepLab的语义阶段在相同的配对RGB图像和一组固定场景的语义映射(i。e.火车场景中每个场景9个)。NeSF可以进一步访问与每个场景的火车摄像机相关联的所有210张RGB地图，这些地图用于适合每个场景的NeRF模型。这两种方法都是通过从新场景中随机抽取的帧来进行评估的，每个场景4个。为了强调NeSF的3D性质，我们评估了来自没有RGB信息的新场景的额外摄像机姿态，从每个场景的测试摄像机中，每个场景有额外的4个姿态。

与稀疏通信网络[44]（3D）的比较。由于稀疏通信网络需要三维输入，我们从摄像机姿态和地面真实深度图中推导出每个场景的预言点云，因此这种方法不公平的优势，在完全三维监督下建立了性能上限。为此，我们使用相同的210个火车帧来拟合NeRF模型。我们进一步选择每个点云的一个子集进行三维语义监督；即与监督NeSF和DeepLab的9个语义映射对应的点。我们在新场景的两组三维点上评估了NeSF和稀疏网络。第一个组是每个点云的子集，对应于从每个场景的训练CAM - ERAS中随机选择的4个帧。这些点可以作为每个场景的3D表示的一部分来使用。第二组是一组额外的查询点，来自每个场景的测试摄像机的4个额外帧。由于稀疏通信网络不是被设计用来对其输入点云之外的点进行分类的，所以我们应用一个最近邻的标签传播过程来为后者分配标签。

定量比较-表2和表3。虽然所有方法在KLEVR数据集上都可以进行比较，但模型质量在更具挑战性的数据集上变化很大。在ToyBox5上，我们的方法的性能与DeepLab相当，但在ToyBox13上，它的性能差6.6%在2D mIoU。虽然我们的方法并没有达到相同的效果作为DeepLab在RGB图像可用的帧上的精度水平，它能够在新的相机姿态上达到接近相同的精度，这是DeepLab无法接近的任务。为了关注我们的方法的基本性质，我们选择将NeSF单独限制为三维几何信息。根据PixelNeRF [142]或IBRNet [136]的精神，通过投影到地面真实摄像机上合并二维信息是很简单的。

正如预期的那样，我们的方法在ToyBox5和19.8-23上的性能也比稀疏通信网络低4.7-5.2%。1%的玩具盒13。与稀疏通信网络不同的是，我们的方法缺乏访问密集的、地面真实的深度地图和完整的3D监督。此外，NeSF所采用的3D UNet体系结构是基于[26]的，这是稀疏通信网络体系结构的前身。由于NeSF没有利用稀疏性，它必须在比基线更低的空间分辨率下运行，并倾向于错误地标记小物体和薄结构。虽然NeSF今天的性能不如稀疏网络，但我们预计模型架构在方法上的改进以快速提高性能。补充材料中包括其他深入分析。

定性比较-图5。定性地说，我们的方法在识别ToyBox13中的13个典型类别方面表现出了很强的性能。因为我们的方法直接运行在3个三维几何上的，它不容易被外观相似但几何不同的物体混淆，就像上面一排的薄步枪所展示的那样。然后

	玩具盒5				玩具盒13			
	训练摄像机		测试摄像机		训练摄像机		测试摄像机	
	2D mIoU	3D mIoU	2D mIoU	3D mIoU	2D mIoU	3D mIoU	2D mIoU	3D mIoU
NeSF	81.9 0.8±	88.7 0.9±	81.7 0.6±	89.6 0.±7	56.5 0.8±	60.1 0.6±	56.6 1.0±	61.9 0.9±
DeepLab [16]	81.6	N/A	N/A	N/A	63.1	N/A	N/A	N/A
SparseConvNet [44]	N/A	93.4	N/A	94.8	N/A	83.2	N/A	81.7

表3。**定量比较-NeSF与二维和三维基线相比具有竞争力。**在训练时，NeSF和DeepLab只使用2D监督。相反，稀疏通信网络需要以标记的3D点云的形式进行完整的3D监督。我们通过反向投影的深度映射来构造神谕点云，从而得到了我们的方法的一个上界（以灰色形式显示的行）。模型被评估在火车和测试摄像机的姿态从测试场景。标记为“N/A”的配置表示方法不适用的设置。NeSF的统计数据通过5个随机初始化进行聚合。

超参数		2D	3D
随机旋转	不	81.1	75.5
	是	92.0	97.1
密度网格	(32, 32, 32)	87.5	92.1
	(48, 48, 48)	91.2	96.0
	(64, 64, 64)	91.7	89.6
	(80, 80, 80)	92.0	97.1
UNet	(16, 32, 64)	89.9	94.4
	(24, 48, 96)	91.5	96.4
	(32, 64, 128)	92.0	97.1
MLP	(0, 32)	91.3	96.4
	(1, 32)	91.8	96.9
	(1, 64)	91.2	96.2
	(2, 128)	92.0	97.1

表4。**消融：超参数-以随机场景旋转的形式增加了数据，提高了密度网格的空间分辨率，增加了UNet模型的容量，提高了二维和三维mIoU。**对来自KLEVR数据集的25个场景进行的实验。

# RGB 图像	NeRF PSNR SSIM				NeSF 2D	3D
5	21:2±	1:4	0:89	0:02±	52:2	72:4
10	24 2± :	1:2	0:92	0:01±	79:6	96:7
25	30 3± :	1:2	0:96	0:01±	87:3	97:0
50	35 5± :	0:9	0:98	0:00±	90:8	97:3
75	37 5± :	1:0	0:98	0:00±	91:4	97:1
100	38 4± :	1:1	0:98 92:0±	0:00		97:4

表5。**消融：对重建质量的敏感度-我们的方法的准确性随着NeRF的重建质量而提高。**PSNR和SSIM在所有场景和聚合的指标中取平均值。对来自KLEVR的所有场景进行的实验。

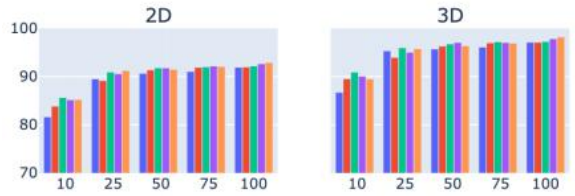

图4。消融：数据效率-2D和3D mIoU作为列车场景数量的函数，每个场景监督1、2、5、10或25个语义映射。NeSF概括到新场景，每个场景只有一个语义映射。每个场景的额外语义映射略微提高了准确性。KLEVR数据集实验。

NeSF和麻雀通信网络正确识别步枪的几何形状，深度实验室标记它与它后面的椅子相同。与DeepLab类似，我们的方法面临着薄结构的挑战，如中间一排的站立灯的管。DeepLab和NeSF分别在二维和三维中使用的常规网格并不能很好地捕捉到这种结构。通过访问密集、准确的点云，稀疏通信网络正确地识别了灯的整体。在NeSF中特别明显的一个限制是倾向于在附近的物体上涂上标签，如底部一行的椅子。*如果无法访问外观或细粒度的几何体，NeSF就无法识别一个对象何时结束，而另一个对象何时开始。*在未来的工作中，整合外观信息和获得更高的空间分辨率是提高NeSF精度的直接方法。

5.2.消融研究

我们的第二组实验研究了我们的方法的每个组成部分如何影响KLEVR数据集上的系统性能。

对特性的敏感性。表4显示了一项消融研究的结果，其中我们的方法的一个特征是不同的，而所有其他方法都是参考值。我们发现，每个组件都提供了一个可测量的改进，在二维和三维分割质量。以随机场景旋转形式进行的数据增强对质量提高最大，2D和3D mIoU分别增加10.3%和11.8%。探测的NeRF密度网格的空间分辨率是第二重要的，因为分辨率不足使较小的物体难以区分。

对重建质量的敏感性-表5。我们研究了NeSF对NeRF重建质量的稳健性。为了调节重建质量，我们改变了重建质量的数量

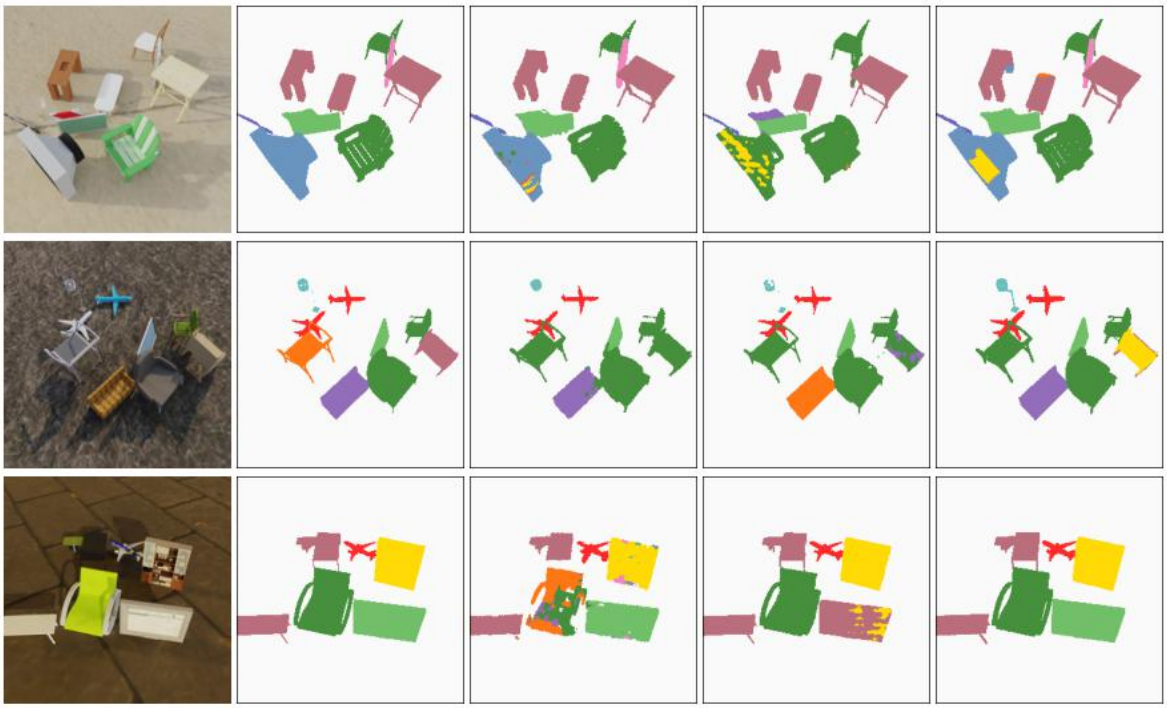

图5。定性比较（ToyBox13）——与DeepLab不同，NeSF能够清晰地分离出外观相似但几何形状不同的物体（顶部）。然而，NeSF很难使用像灯柱（中间）这样的薄结构，并倾向于从附近的物体上涂抹标签（底部）。稀疏通信网络没有任何限制，但可以访问oracle 3D几何和完整的3D监督。

当拟合从5到100的NeRF模型时使用的RGB图像。正如预期的那样，随着提供更多的RGB图像，在新视图上测量的NeRF重建质量会增加。同时，我们发现NeSF二维分割精度随着NeRF重建质量的提高而单调地提高。提高NeRF重建质量是一个高度活跃的研究领域，因此我们预计方法上的改进可以直接应用于提高NeSF的性能。令人惊讶的是，当NeRF模型仅用25张RGB图像进行优化时，3D分割精度达到接近97%。

为了研究NeSF在标记语义映射稀缺的场景中的适用性，我们研究了对每个场景的语义映射数量的鲁棒性。我们发现NeSF可以很容易地推广到新的场景中，每个火车场景只需一个语义映射。对于少量的场景，额外的语义映射每个场景可以提高性能，在25个场景后没有明显的效果。这表明由视频组成的数据集，每个都有一个标记帧，是NeSF的理想选择。

6. 结论和局限性

在这项工作中，我们提出了NeSF，一种新的方法，同时进行三维场景重建和语义分割从摆出的二维图像。在NeRF的基础上，我们的方法仅在姿态的二维RGB图像和语义映射上进行训练。在推理时，我们的方法构建了一个密集的语义分割域，可以直接在三维图像中进行查询，也可以用于从新的摄像机姿态中呈现二维语义映射。

我们在三个新的数据集上比较了NeSF与二维和三维语义分割的竞争基线。

在更具挑战性的环境中，我们发现NeSF的表现低于其基线。然而，NeSF提供了新的功能。与传统的二维分割方法不同，NeSF通过多个独立的视图融合信息，并从新的姿态呈现语义映射。与3D点云方法不同，NeSF在训练和测试时间都单独操作提出的2D信息。我们选择将NeSF限制在一组核心特性上，以更好地探索这种方法的基本权衡和能力。我们在补充材料中进一步探讨了模型选择的权衡和潜在的社会影响。在未来的工作中，我们预计将NeSF扩展到包含二维语义模型和三维稀疏性将显著提高准确性。

除了NeSF之外，我们还提出了三个新的数据集，用于多视点三维重建和语义分割，总计超过300万帧和1000个场景。每个数据集包含数百个场景，每个场景都由一组随机放置的对象组成。这些都更具挑战性

数据集是呈现与真实的照明和大量的对象和背景目录。这些数据集以及伴随的代码和预先训练过的NeRF模型，将在发布后向公众发布。

致谢

我们想对康斯坦丁诺斯·雷马塔斯表示感谢。斯卡利，特别是托马斯·芬克豪瑟的想法和建议。我们要注意的是，我们对雅各科普利公司提供的项目支持和领导深表赞赏，没有这些，就不可能工作。作者还要感谢蔡洛基协助组装深度实验室基线实验。

参考文献

[1] Humam·阿尔瓦塞尔、德鲁夫·马哈詹、布鲁诺·科尔巴尔、洛伦佐·托雷萨尼、伯纳德·哈纳姆和杜·特兰。通过跨模态音频-视频聚类的自监督学习。*神经信息处理系统的研究进展，2020年33日*。2

[2]，亚美尼亚，何志洋，荣永，阿米尔。扎米尔马丁·费舍尔，吉坦德拉·马利克和西尔维奥·萨瓦雷斯。三维场景图：一种用于统一语义、三维空间和照相机的结构。*在2019年IEEE国际计算机视觉会议论文集上*。2

[3] Yusuf Aytar，卡尔·冯德里克和安东尼奥·托拉尔巴。声网：从未标记的视频中学习声音表示。*在神经信息处理系统的进展中，第892-900页，2016年*。2

[4] Vijay巴德里纳拉亚南，亚历克斯·肯德尔和罗伯托·西波拉。Segnet：一种用于图像分割的深度卷积编解码器架构。IEEE关于模式分析和机器智能的交易，39（12）：2481-2495,2017。2

[5]，乔纳森·巴伦，本·米尔登霍尔，马修·坦西克，彼得海德曼，里卡多马丁-布鲁拉，和斯里尼瓦桑。一种抗锯齿神经辐射场的多尺度表示法。在ICCV，2021年10月。3

[6]奥德·比拉德和丹妮卡·克拉吉奇。机器人操作的趋势和挑战。科学，364（6446），2019。1

[7]弗朗西斯科波宁-字体，阿尔贝托奥尔蒂斯，加布里埃尔奥利弗。针对移动机器人的视觉导航：一项调查。*杂志的*智能和机器人系统，53(3)：263-296,2008年。1 [8]马克老板，拉斐尔布劳恩，瓦伦贾帕尼，乔纳森巴尔-龙，刘赛德，和亨德里克·伦斯奇。从图像集合中得到的神经反射率分解。在ICCV，2021年10月。3

[9] Holger凯撒，瓦伦·班基蒂，亚历克斯·朗，苏拉布·沃拉，威尼斯刘爱玲、徐强、南、于潘、巴尔丹、奥斯卡北鲍。一个用于自动驾驶的多模态数据集。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第11621-11631页，2020年*。1

[10] Holger凯撒，瓦伦·班基蒂，亚历克斯·朗，苏拉布·沃拉，威尼斯刘玲、徐强、余潘、吉安卡罗·巴尔丹和奥斯卡·北京宝贝。一个用于自动驾驶的多模态数据集。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第11621-11631页，2020年*。2，5

[11] Rohan Chabra，简·伦森，艾迪·伊尔格，坦纳·施密特，朱-利安·斯特劳布，史蒂文·洛夫格罗夫和理查德·纽科姆。深度局部形状：学习局部SDF先验，以进行详细的三维重建。在ECCV，2020年。3

张天使，戴安吉拉，托马斯·芬克舍瑟，马西杰哈尔伯、马提亚斯尼斯纳、马诺利斯萨瓦、宋秀兰、曾安迪和张银达。在室内环境中学习rgb-d数据。*arXiv预印本：1709,2017。.06158*2，5

[13]天使X Chang，托马斯·芬克豪瑟，列奥尼达斯·吉巴斯，韩拉罕、黄七星、李齐莫、萨瓦、宋秀兰、苏浩等。一个信息丰富的三维模型存储库。*arXiv预印本，arXiv：1512.03012,2015年*。5

[14]厄卓车，荣格和奥尔森。移动激光扫描点云的目标识别、分割和分类：一个最新的综述。传感器，19(4)：810,2019。2

陈安培和徐泽翔。MVSNeRF：快速将军能够从多视点立体声音响中进行辐射场重建。在ICCV，2021年10月。3

[16] Liang-Chieh陈，乔治·帕潘安德里奥，柯基诺斯，凯文墨菲和艾伦L尤尔。Deeplab：具有深度卷积网、无对称卷积和全连通crfs的语义图像分割。IEEE关于模式分析和机器智能的交易，40(4)：834-848,2017。2，6，7，8

[17] Liang-Chieh陈，乔治·帕潘德里欧，弗洛里安·施罗夫，和哈特维格亚当。语义图像分割的重构式卷积。*arXiv预印arXiv：1706.05587，*2017.2

陈[18]志勤、安德烈、张浩。BSP-网络：通过二元空间划分生成紧凑的网格。在CVPR，2020年。3

陈[19]志勤、张浩。学习生成式形状建模的隐式领域。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第5939-5948页，2019页*。2

陈[20]志勤、张浩。学习隐式字段生成形状建模。在CVPR，2019年。3

[21]跑成，瑞安拉扎尼，以山塔哈维，李和刘冰冰。（af）2-s3net：稀疏语义分割网络采用自适应特征选择的注意特征融合。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第12547-12556页，2021页*。2

[22]Franc,ois球。具有深度可分离卷积的深度学习。*发表在IEEE计算机视觉和模式识别会议论文集上，第1251-1258页，2017年*。2

[23]克里斯托弗·乔，荣扬·格瓦克和西尔维奥·萨瓦雷斯。四维时空对流网络：闵可夫斯基卷积神经网络。*在IEEE/CVF会议会议记录中*

*《计算机视觉和模式识别》，第3075-3084页，2019年。*1

[24]克里斯托弗·乔，荣扬·格瓦克和西尔维奥·萨瓦雷斯。四维时空对流网络：闵可夫斯基卷积神经网络。*在IEEE关于计算机视觉和模式识别的会议的论文集中，第3075-3084页，*

2019.2

[25]克里斯托弗·乔伊，杰西克·帕克，和弗拉德伦·科尔顿。完全卷积的几何特征。*发表在IEEE国际计算机视觉国际会议论文集，第8958-8966页，2019页。*2

[26] O¨ 还有：艾哈迈德·阿卜杜勒卡迪尔、索伦·利恩坎普、托马斯·布罗克斯和奥拉夫·朗内伯格。3d u-net：从稀疏注释中学习密集的体积分割。在*医学图像计算和计算机辅助干预问题国际会议，第424-432页。*施普林格，2016年。4，6，15

[27]福雷斯特·科尔，凯尔·热诺娃，大卫，丹尼尔·弗拉西奇，张智通。通过不可微采样进行可微表面渲染。在ICCV，第6088-6097页，

2021.3

[28]搅拌机在线社区。*搅拌机-一个三维建模和渲染软件包。*搅拌机基金会，固定搅拌机基金会，阿姆斯特丹，2018年。5

[29] Tiago·科廷哈尔，乔治·齐勒皮斯和埃伦·埃尔达尔·阿克索伊。用于自动驾驶的激光雷达点云的快速、不确定性感知的语义分割。*arXiv预印本：2003,2020。.036532*

[30]安吉拉，戴，天使X张，马诺利斯萨瓦，马西吉哈尔-伯，托马斯·芬克豪舍尔和马提亚斯·Nießner。注释丰富的室内场景的三维重建。在*IEEE计算机视觉和模式识别会议论文集，第5828-5839页，2017年。*1，2，5

[31]·安吉拉·Dai和马提亚斯·Nießner。3dmv：联合3d-多三维语义场景分割的视图预测。在*欧洲计算机视觉会议（ECCV）论文集，第452-468页，2018页。*1，2

[32]·安吉拉·戴，丹尼尔·里奇，马丁·博克洛，斯科特·里德，Jrgen Sturm，和马提亚斯·Nießner。扫描完成：三维扫描的大尺度场景完成和语义分割。*发表在IEEE计算机视觉和模式识别会议论文集上，第4578-4587页，2018页。*3

[33]博阳，凯尔热诺娃，布阿齐兹，杰弗里欣顿，安德里亚·塔利亚萨奇和索罗什·亚兹达尼。可学习的凸凸分解。*在CVPR，2020年。*3，5

[34]博阳邓，JP刘易斯，蒂莫西耶鲁扎尔斯基，杰拉德庞斯-摩尔、杰弗里·辛顿、穆罕默德·诺鲁齐和安德里亚·塔利亚萨奇。NASA：神经铰接式的形状近似值。在ECCV，2020年。3

[35]特伦斯DeVries。无约束场景生成局部条件的辐射场。在ICCV，2021年10月。3

[36]伯特兰·杜伊拉德，詹姆斯·安德伍德，诺亚·昆茨，弗拉斯金，阿拉斯泰尔夸德罗斯，彼得莫顿，和阿隆弗兰克尔。关于三维激光雷达点云的分割。*2011年IEEE机器人与自动化国际会议，第2798-2805页。*2011年IEEE。1

[37] Ariel埃弗拉特，因巴尔·莫塞里，奥兰·朗，塔利·德克尔，凯文

威尔逊，阿维纳坦·哈西迪姆，威廉·T。弗里曼和迈克尔·鲁宾斯坦。希望在鸡尾酒会上收听：一个独立于演讲者的演讲视听模型

定量*ACM跨。图*，37(4)，2018年7月。2

[38]庄甘，杭昭，陈培浩，大卫考克斯，安东尼奥托拉尔巴。自我监督的移动车辆跟踪与立体声。*发表在IEEE国际计算机视觉会议论文集，第7053-7062页，2019页。*2

[39]·斯蒂芬·加尔宾和马雷克·科瓦尔斯基。FastNeRF：高保真度神经渲染在200FPS。在ICCV，2021年10月。3

[40]凯尔·热诺娃，弗雷斯特·科尔，丹尼尔·弗拉西奇，亚伦·萨尔纳，

威廉·弗里曼和托马斯·芬克豪瑟。使用结构化的隐式函数来学习形状模板。在ICCV，2019年。3，5

[41] Kyle热诺娃，尹小七，昆都，卡罗琳·潘托-法鲁，弗雷斯特科尔，大卫，布莱恩布雷温顿，布莱恩肖克，和托马斯芬克豪瑟。学习三维语义分割，只有二维图像监督。3DV，2021.1，2，16

[42] Rohit吉尔达，杜特兰，洛伦佐托雷萨尼和拉曼南。分散注意力：学习没有单一标记视频的视频表示。*在IEEE计算机视觉国际会议论文集，第852-861页，*

2019.2

[43]本杰明格雷厄姆，马丁恩格尔克，和劳伦斯范德Matat。基于子流形稀疏卷积网络的三维语义分割。*发表在IEEE计算机视觉和模式识别会议论文集上，第9224-9232页，2018年。*2，16

[44]本杰明·格雷厄姆和劳伦斯·范德马顿。子流形稀疏卷积网络。*arXiv预印本，arXiv：1706.01307,2017年。*6，7，8

[45] Klaus Greff和安德里亚·塔利亚萨奇。库布里克。https:

//github.com/google-research/kubric，2021.5，15

[46]蒂博集团，马修·费舍尔，弗拉基米尔·G。基姆布莱恩C。拉塞尔和马修·奥布里。â€一种学习三维表面生成的论文研究方法。*在CVPR，2018年。*5

[47]郭玉兰、王汉云、胡庆勇、刘浩、刘李、和穆罕默德·本纳蒙。对三维点云的深度学习：一项调查。*IEEE交易上的模式分析和机器智能，2020年。*2

[48]Sauraph古普塔，朱迪·霍夫曼和吉坦德拉·马利克。采用跨模态蒸馏法进行监督转移。*发表在IEEE计算机视觉和模式识别会议论文集，第2827-2836页，2016页。*2

[49]阿卜杜勒·哈菲兹和古拉姆·穆希丁兄弟。关于实例细分的调查：最新水平。*国际多媒体信息检索期刊，第1-19页，*

2020.1

[50]雷韩、田正、兰旭、陆方。奥克塞格：可实现占用率感知的三维实例分割。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第2940-2949页，2020页。*2

[51] Rana • 哈诺卡，阿米尔 • 赫兹，诺亚鱼，拉贾 • 吉里耶斯，沙查尔
弗莱什曼和丹尼尔 • 科恩。一个有优势的网络。*ACM图形交易（TOG），38(4)：2019年1-12日*。1，2

何[52]，余红山、刘晓燕、杨正英、魏
孙、王雅南、傅强、邹艳梅、阿雅马尔棉。基于深度学习的三维分割：一项调查。*arXiv预印本arXiv：2103.05423,2021年*。2

[53]，彼得 • 海德曼，普拉图尔 • 斯里尼瓦桑，本 • 米尔登霍尔，乔纳森
巴伦和保罗 • 德贝维克。用于实时视图合成的烘焙神经辐射场。在ICCV，2021年10月。3

[54]菲利普亨兹勒。三维对象计算机的无监督学习
来自野外的视频。在CVPR，2021年。3

[55]亚历山大 • 赫曼斯，乔治斯 • 弗洛洛斯和巴斯蒂安 • 莱比。从rgb-d图像中进行室内场景的密集三维语义映射。*2014年IEEE机器人与自动化国际会议（ICRA），第2631-2638页*。2014年IEEE。2

[56]胡泽宇、甄明明、白旭阳、傅洪波、和
齐伦泰。Jsenet：针对三维点云的联合语义分割和边缘检测网络。在ECCV，2020年。
2

[57]黄精卫、张浩天、李毅、方克豪泽，
马提亚斯 • Nießner和列奥尼达斯 • 吉巴斯。文本集：从网格上的高分辨率信号中学习的一致性局部参数化。*在IEEE计算机视觉和模式识别会议的论文集上，*
第4440-4449页，2019年。2

[58]芮黄，张万岳，昆都，卡罗琳潘
法鲁，大卫A罗斯，托马斯芬克豪瑟，和阿里雷萨法蒂。一种在激光雷达点云中进行时间三维目标检测的lstm方法。*在计算机视觉-ECCV2020：第16届欧洲会议，格拉斯哥，英国，2020年8月23-28日，论文集，第十八部分，第16页，第266-282页*。施普林格，2020年。
16

[59] Shahram Izadi，大卫 • 金，奥特马尔 • 希里格斯，大卫莫利诺，理查德 • 纽科姆，普什米特 • 科利，杰米 • 肖顿，史蒂夫 • 霍奇斯，达斯汀 • 弗里曼，安德鲁 • 戴维森等人。运动融合：使用移动深度摄像机进行实时三维重建和交互。*发表在第24届ACM用户界面软件与技术年度研讨会论文集中，第559-568页，2011年*。1

[60] Ajay Jain，马修 • 坦西克和彼得 • 艾比尔。放
饮食上的NeRF：语义上一致的少镜头视图合成。在ICCV，2021年10月。3

张邦[61]和卢尔德阿加皮托。CodeNeRF：已删除
物体类别的纠缠神经辐射场。在ICCV，2021年10月。3

[62]公司，顾佳缘，郝苏。多视图
点网为3d场景理解。*发表在IEEE计算机视觉研讨会国际会议论文集，第0-0页，2019年*。2

[63] Yoonwoo Jeong。自校准的神经辐射场。
在ICCV，2021年10月。3

[64]金义伟，江二琼、蔡明。使用深度学习的三维重建：一项调查。*《信息与系统中的通信》，20(4)：389-413年，2020年*。2

[65]龙龙静、陈玉成、张玲、何明义、英利天。自监督模态和视图不变特征学习。*arXiv预印本arXiv：2005.14169,2020年*。2

[66]，贾斯汀 • 约翰逊，巴拉斯 • 哈里哈兰，劳伦斯 • 范德 • 马顿，
李飞飞，C劳伦斯齐特尼克和罗斯吉希克。Clevr：一个用于组合语言和基本视觉推理的诊断数据集。*发表在IEEE计算机视觉和模式识别会议论文集上，第2901-2910页，2017年*。5

[67]饮食协会，金马和吉米 • 巴。一种随机优化的方法。*arXiv预印arXiv：1412.6980，*
2014.15

[68]是索菲亚 • 科普克，奥利维亚 • 怀尔斯，耶埃尔 • 摩西，安德鲁 • 齐塞尔曼。视觉到声音：视觉钢琴抄写的端到端的方法。*在ICASSP 2020-2020 IEEE声学、语音和信号处理国际会议（ICASSP）上，第1838-1842页*。2020年IEEE。2

[69]布鲁诺 • 科尔巴，杜特兰和洛伦佐 • 托雷萨尼。来自自监督同步的音频和视频模型的协同学习。*在神经信息处理系统的进展中，第7763-7774页，2018年*。2

[70]亚当 • 科西奥雷克，海科 • 斯特拉斯曼，丹尼尔 • 佐兰，波尔 • 莫雷诺，罗莎莉亚 • 施耐德，索纳 • 莫克尔和丹尼洛 • 雷曾德。NeRF-VAE：一个具有几何感知的三维场景生成模型。在ICML，2021年。3

[71] Adarsh • 科德尔，克里斯托弗 • 雷曼，肖恩 • 法内洛，安德里亚
塔利亚萨、乔纳森泰勒、菲利普戴维森、窦明松、郭开文、森凯斯金、卡米斯等。需要4个速度在实时密集的视觉跟踪。*ACM图形交易（TOG），37(6)：2018年1-14日*。1

[72]亚历克斯 • 克里热夫斯基，伊利亚 • 苏茨克弗和杰弗里 • 辛顿。
基于深度卷积神经网络的图像集分类。*神经信息处理系统的进展，25：1097-1105,2012*。1

[73]Kundu，尹小七，阿里丽莎法蒂，大卫罗斯，布莱恩
布鲁温顿、托马斯 • 芬克豪瑟和卡罗琳 • 潘托法鲁。虚拟多视图融合技术用于三维语义分割。*在欧洲计算机视觉会议上，第518-535页*。施普林格，2020年。1，2

[74] K.赖，L。博和D。狐狸无监督特征学习
用于3d场景标签。*2014年IEEE机器人与自动化国际会议（ICRA），第3050-3057页，2014年*。2

[75] Fahad • 拉蒂夫和亚辛 • 鲁切克。使用深度学习技术的语义分割的调查。*Neurocomputing, 338:321 - 348, 2019.*1

[76] Felix • 杰雷莫 • 拉文，马丁 • 丹内尔詹，帕特里克 • 托斯特伯格，
古塔姆 • 巴特，法哈德 • 沙巴兹汗和迈克尔 • 费尔斯伯格。深度投影三维语义分割。*在图像和模式的计算机分析国际会议上，第95-107页*。施普林格，2017年。2

林陈贤，马伟九，托拉巴和斯-
蒙露西。神经束调整神经辐射场。在ICCV，2021年10月。3

[78]宗-林毅、迈克尔 • 梅尔、塞尔日 • 贝隆吉、詹姆斯 • 海斯、彼得罗纳、拉曼南、彼得多尔和劳伦斯 • 齐特尼克。微软coco：上下文中的常见对象。在

欧洲计算机视觉会议，第740-755页。施普林格，2014年。1

[79]，大卫·林德尔，朱利安·马特尔，和戈登·韦茨斯坦。Au-快速神经卷渲染的自动集成。在CVPR，2021年。3

刘[80]灵杰、顾嘉道、林觉、田当苏、基督教玄武岩。神经稀疏体素场。*在Adv。神经信息。过程。西斯特。*，2020.2，3，4

刘李[81]、欧阳万里、王小刚、菲格、陈杰、刘新旺、马蒂·皮提基宁。针对通用对象检测的深度学习：一项调查。国际计算机视觉杂志，128(2)：261-318,2020年。1

刘卫平、孙贾、李万一、胡婷、王鹏。点云及其上的深度学习和应用：调查。*传感器，19（19）：4188,2019*。2

[83]，乔纳森·朗，埃文·谢尔哈默，和特雷弗·达雷尔。用于语义分割的全卷积网络。*发表在IEEE计算机视觉和模式识别会议论文集上，第3431-3440页，2015页*。2

[84] Lingni Ma，Jrg·斯特克勒，克里斯蒂安·克尔和丹尼尔·克雷默斯。多视图深度学习，以实现一致的语义映射与rgb-d相机。*2017年，IEEE/RSJ智能机器人与系统国际会议（IROS），第598-605页*。2017年IEEE。2

[85]里卡多·马丁-布鲁拉拉，诺哈·拉德万，梅赫迪·萨贾迪，乔纳森·巴伦，阿列克谢·多索维茨基和丹尼尔·达克沃斯。野外的NeRF：不受约束的照片收集的神经辐射场。在CVPR，2021年。3

[86]·鲁本·马斯卡罗，卢卡斯·特谢拉和玛格丽塔·切利。扩散器：多视角二维到三维标签扩散的语义场景分割。*在IEEE机器人和自动化国际会议（ICRA2021）（虚拟），2021年*。2

[87]乔纳森·马西，大卫·博斯卡尼，迈克尔·布朗斯坦，和皮埃尔范德金斯特。黎曼流形上的测地线卷积神经网络。*IEEE计算机视觉研讨会国际会议记录，第37-45页*。1

[88]纳尔逊马克斯。用于直接体积渲染的光学模型。*IEEE可视化和计算机图形学学报，1(2)：99-108,1995*。3

[89]约翰麦科马克，安库尔汉达，安德鲁戴维森，和Ste-风扇Letenger。语义融合：使用卷积神经网络的密集三维语义映射。*2017年IEEE机器人与自动化国际会议（ICRA），第4628-4635页*。2017年IEEE。2

[90] Quan孟。基于gan的神经辐射场没有摆好的相机。在ICCV，2021年10月。3

[91] Lars梅切德，迈克尔奥克斯尔，迈克尔尼迈耶，Se-巴斯蒂安·诺沃津和安德烈亚斯·盖格。占用网络：在功能空间中学习三维重建。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第4460-4470页，2019年*。2

[92] Lars梅切德，迈克尔奥克斯尔，迈克尔尼迈耶，Se-巴斯蒂安·诺沃津和安德烈亚斯·盖格。占用网络：学习函数空间中的三维重建。*在CVPR，2019年*。3，5

[93]，约翰内斯·迈耶，安德烈亚斯·艾特尔，托马斯·布罗克斯，和沃尔-弗拉姆伯加德。利用多模态对比学习改进单峰目标识别。*在IEEE/RSJ智能机器人和系统国际会议上，2020年*。2

[94]，本·米尔登霍尔，斯里尼瓦桑，马修·坦西克，乔纳森巴伦，拉维拉莫莫蒂，和任吴。NeRF：将场景表示为视图合成的神经辐射场。在ECCV中，第405-421页。施普林格，2020年。2，3，4，5，15，16

[95] A.米利奥托，我。Vizzo，J.贝利和C。斯塔奇尼斯。RangeNet++：快速和准确的激光雷达语义分割。*在IEEE/RSJ国际。会议关于智能机器人和系统（IROS），2019年*。2

[96]谢尔文·米奈，尤里·博伊科夫，法蒂赫·波里克利，安东尼奥·J广场，纳赛尔凯塔纳瓦兹和德米特里特佐普洛斯。使用深度学习的图像分割：一个调查。*《IEEE《模式分析与机器智能学报》，2021年*。2

[97] Zak Murez，塔伦斯·范·阿斯，詹姆斯·巴托洛齐，阿扬·辛哈，维贾伊·巴德里纳拉亚南和安德鲁·拉比诺维奇。地图集：从姿态图像的末端三维场景重建。*在ECCV，2020年*。2，6

[98] Arsha Nagrani，陈孙，大卫·罗斯，拉胡尔·苏克坦卡尔，科迪莉亚·施密德和安德鲁·齐瑟曼。语音2动作：动作识别的跨模态监督。*在2020年IEEE计算机视觉和模式识别会议论文集上*。2

[99]托马斯Neff。DONeRF：面向实时渲染利用深度Oracle网络研究的紧凑神经辐射场。*在欧洲图形，2021年*。3

[100]的迈克尔·尼迈耶和安德烈亚斯·盖格。长颈鹿：发送场景作为组成生成的神经特征域。在CVPR，2021年。3

[101]，迈克尔·尼迈耶，拉尔斯·梅切德，迈克尔·奥克斯尔，和安德烈亚斯盖格。可微体积渲染：在没有三维监督的情况下学习隐式三维表示。*在CVPR，2020年*。3

[102]Noguchi宏。神经关节辐射场。在ICCV，2021年10月。3

[103]，迈克尔·奥克斯勒，宋友鹏，和安德烈亚斯·盖格。单神经元：统一的神经隐式表面和辐射场，用于多视图重建。在ICCV，2021年10月。3

[104]安德鲁·欧文斯，吴家俊，何希H。麦克德莫特，威廉T。弗里曼和安东尼奥·托拉尔巴。环境声音为视觉学习提供监督。*发表在2016年欧洲计算机视觉会议（ECCV）的论文集上*。2

[105]郑琼公园，彼得·弗罗伦斯，朱利安·斯特劳布，理查德纽科姆和史蒂文·洛夫格罗夫。Deepsdf：学习连续有符号的距离函数的形状表示。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第165-174页，2019年*。2

[106] Jeong琼公园，皮特佛罗伦斯，朱利安斯特劳布，理查德纽-

和史蒂文·洛夫格罗夫。DeepSDF：学习连续符号距离函数。
在CVPR，2019年。3

[107]昆洪公园，乌特卡什·辛哈，乔纳森·巴伦，索菲恩
布阿齐兹，丹·戈德曼，史蒂文·塞茨和里卡多·马丁布鲁阿拉。神经化：可变形的神经辐射场。在ICCV，2021年10月。3

[108] Sida彭。神经体：隐式神经表征与结构化的潜在代码，用于动态人类的新观点合成。在CVPR，2021年。3

[109]，迈克尔·尼迈耶，拉尔斯·梅切德，马克
波莱费斯和安德烈亚斯盖格。卷积占用网络。*在欧洲计算机视觉会议（ECCV）上，章，8月*。2020.施普林格国际出版公司。3

[110]广辉，阮丹，华平子，罗丽杰和杨世基。Jsis3d：基于多任务点态网络和多值条件随机场的三维点云的联合语义实例分割。在
IEEE计算机视觉和模式识别会议论文集，第8827-8836页，2019页。2

[111]杰拉德庞斯莫尔。在特征空间中的隐式函数
三维形状重建和完成。在CVPR，2020年。
3

[112]齐、郝苏、莫开春、列奥尼达·吉巴斯。
点网：对点集进行深度学习，以进行三维分类和分割。*发表在IEEE计算机视觉和模式识别会议论文集上，第652-660页，2017年*。1，2，3，4

[113]是齐国、李毅、郝苏、列奥尼达·J
圭巴斯。点网++：在度量空间中对点集的深度层次特征学习。*《神经信息处理系统的进展》，第5099-5108页，2017年*。2，
4

[114]丹尼尔·瑞本，魏江，雅兹达尼，柯李，
光木毅和塔利亚志。分解的辐射场。
https://arxiv.org/abs/2011.12490，
2020.2

[115]教雷泽，彭松友，廖依依，安德烈亚斯
盖革KiloNeRF：用数千个微小的mlp来加速神经辐射场。在ICCV，2021年10月。3

[116]康斯坦丁诺斯雷马塔斯，里卡多马丁-布鲁拉拉，和维托里奥法拉利ShaRF：来自单一视图的形状条件辐射场。*在ICML，2021年*。3

[117] Gernot·里格勒，阿里·奥斯曼·乌鲁索伊，和安德烈亚斯·盖格尔。
Octnet：学习高分辨率的深度三维表示。*发表在IEEE计算机视觉和模式识别会议论文集上，第3577-3586页，2017年*。2

[118]奥拉夫·朗内伯格，菲利普·菲舍尔和托马斯·布罗克斯。U-网络：用于生物医学图像分割的卷积网络。*在医学图像计算和计算机辅助干预国际会议上，第234-241页*。施普林格，2015年。1

[119]奥拉夫·朗内伯格，菲利普·菲舍尔和托马斯·布罗克斯。U-网络：用于生物医学图像分割的卷积网络。*在医学图像计算和计算机辅助干预国际会议上，第234-241页*。施普林格，2015年。2

齐藤顺助，曾黄，夏田，茂茂
三岛、金泽角珠和李浩。PIFu：像素对齐
高分辨率服装人体数字化的隐式功能。在ICCV，2019年10月。3

[121] Katja·施瓦茨，廖依依，迈克尔·尼迈耶和安德烈亚斯盖革用于3d感知图像合成的生成辐射场。*在Adv。神经信息。过程。西斯特。*，2020.3

[122]石少帅、郭卓、李江、王哲、建平
石氏、王小刚、李宏生。PV-RCNN：用于三维目标检测的点体素特征集抽象。在
IEEE/CVF计算机视觉和模式识别会议论文集，第10529-10538页，2020页。
2

[123]舒兰宋，余费雪，曾安迪，X张天使，马野
莉斯·萨瓦和托马斯·芬克豪瑟。语义从单一深度图像的场景完成。*发表在IEEE计算机视觉和模式识别会议论文集上，第1746-1754页，2017年*。2

[124]出版社，邓博阳出版社，张秀明出版社，马修出版社
坦西克，本·米尔登霍尔和乔纳森·巴伦。NeRV：用于重新和视图合成的神经反射和可见场。在CVPR，2021年
。3

[125]朱利安·斯特劳布，托马斯·惠兰，马林尼，陈玉凡，
埃里克维曼，西蒙格林，雅各布J恩格尔，劳尔穆尔塔尔，卡尔伦，肖希特维尔马，等。复制数据集：室内空间的数字副本。*arXiv预印本：1906，.05797*
2019.2

[126]埃德加·苏卡尔，刘世坤，约瑟夫·奥尔蒂斯和安德鲁·戴维森。
iMAP：实时的隐式映射和定位。在ICCV，2021年10月。3

[127] Pei孙，亨里克·克雷兹施马，薛西斯·多蒂瓦拉，奥雷林
周阿德、帕特奈克、徐克、郭建美、周尹、柴玉宁、凯恩等。自动驾驶感知的可扩展性： Waymo开放数据集。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第2446-2454页，2020年*。5

[128]塔吉川。神经几何层次的细节：真实的
使用隐式三维形状的时间渲染。在CVPR，2021年。2，3

[129] Hugues托马斯，查尔斯·齐，让-伊曼纽尔·德肖，比奥特丽斯·马科特吉，古莱特和列奥尼达·吉巴斯。对于点云的灵活和可变形的卷积。*发表在IEEE国际计算机视觉国际会议论文集，第6411-6420页，2019页*。2

[130]永隆田，克里希南，伊索拉。对比多视图编码。*计算机视觉-ECCV2020：第16届欧洲会议，格拉斯哥，英国，2020年8月23-28日，会议记录，第11-16部分，第776-794页*。施普林格，2020年。2

[131] Du Tran，卢布米尔·布尔德夫，罗伯·费格斯，洛伦佐·托雷萨尼，
和马诺哈尔帕卢里。利用三维卷积网络学习时空特征。*《IEEE计算机视觉国际会议论文集》，第4489-4497页，2015页*。1

[132]埃德加·特雷茨克，阿尤什·特瓦里，弗拉迪斯拉夫·戈利亚尼克，迈克尔·佐尔费弗，克里斯托夫·拉斯纳和克里斯蒂安·西奥博特。非刚性神经辐射场：单眼视频中变形场景的重建与新视图合成。在ICCV，2021年10月。3

[133]维尼特，米克西克，莫滕·利德加德，马提亚斯·Nießner，斯图尔特·戈洛德茨，维克多·普里萨卡留，奥拉夫·科勒，大卫·默里，沙拉姆·伊扎迪，帕特里克·普雷兹等。增量密集语义立体融合用于大规模语义场景重建。*2015年IEEE机器人与自动化国际会议（ICRA），第75-82页*。2015年IEEE。2

[134] Sourabh Vora，亚历克斯·H·朗，巴萨姆·海鲁，和奥斯卡北京博姆。点绘制：三维目标检测的顺序融合。*在IEEE/CVF计算机视觉和模式识别会议的论文集中，第4604-4612页，*2020.2

王海燕、荣学健、梁杨、王水华、和英利田。针对野生场景的三维图结构点云中的弱监督语义分割。在BMVC，第284页，2019年。2

王[136]倩倩，王志诚，热诺娃，斯里尼-瓦桑、周霍华德、乔纳森·巴伦、里卡多·马丁布鲁拉、诺亚·斯纳弗利和托马斯·芬克豪瑟。IBRNet：学习基于多视图图像的渲染。在CVPR，2021年。2，3，6

[137]王智瑞，吴尚哲，谢伟迪，陈敏，还有维克多·阿德里安·普里萨卡里留。神经神经没有已知相机参数的辐射场。*https://arxiv.org/abs/2102.07064*，2021.3

吴紫峰，沈春华，范登恒尔。更宽或更深的：重新访问resnet模型的可视化认识*模式识别，2019年90：119-133*。6,16 [139] Lior Yariv先生，尤尼·卡斯滕先生，德罗·莫兰先生，梅拉夫·加伦先生，马坦先生

阿茨蒙，罗南·巴斯里和雅隆·利普曼。通过分离几何形状和外观而进行的多视图神经表面重建。*在Adv。神经信息。过程。西斯特。*，2020.3

[140]林延陈，皮特·弗洛伦斯，乔纳森·巴伦，阿尔贝托·罗-德里格斯，菲利普·伊索拉，和林宗义。用于姿态估计的反转神经辐射场。在IROS，2021年。3

[141]余历、李瑞龙、天奇、李浩、任、Kanazawa江。神经辐射场实时渲染的平面八树。在ICCV，2021年10月。3，18

[142]亚历克斯，叶薇姬，马修坦西克和金泽安乔。来自一个或几张图像中的神经辐射场。在CVPR，2021年。2，3，6

[143]、袁文涛、吕昭阳、坦纳·施密特、史蒂文洛夫格罗夫住所名称基于神经渲染的运动中刚性物体的自监督跟踪和重建。*在CVPR，2021年*。3

[144] Greg Zaal，罗布图伊特尔，里科西里尔，詹姆斯雷科克，安-米乔克、马吉博罗达、迪米特里奥斯萨瓦和朱里塔汉堡。Hdri天堂。hhttps://polyhaven .com/ hdris，2021.5

张[145]、刘志、刘广文、黄丹东。使用单目视觉的大规模三维语义映射。*2019年IEEE第四届图像、视觉和计算国际会议（ICIVC），第71-76页*。IEEE，2019。2

[146]张秀明，邓博阳，保罗德贝维克，威廉T弗里曼，和乔纳森T巴伦。神经因子：条件下形状和反射率的神经因子分解未知的照明。*arXiv预印arXiv：2106.01970，*2021.2

张[147]，周子祥，腓力大卫，悦翔宇，泽荣西、龚庆、福桑。偏振网：一种改进的在线激光雷达点云语义分割的网格表示方法。*发表在IEEE/CVF计算机视觉和模式识别会议论文集上，第9601-9610页，2020页*。2

[148]杭赵，庄甘，安德鲁鲁琴科，卡尔冯-德里克，乔什·麦克德莫特和安东尼奥·托拉尔巴。像素的声音。*请参见《欧洲计算机视觉会议（ECCV）的会议记录》，第570-586页，2018年*。2

[149]，特里斯坦莱德洛，斯特凡，和安德鲁戴维森。用隐式场景表示来标记和理解就地场景标签。在ICCV，2021年10月。2，3，4

朱[150]、周慧、大王、洪方舟、悦信马、李伟、李宏生、林大华。用于激光雷达分割的圆柱形和不对称三维卷积网络。*在IEEE/CVF计算机视觉和模式识别会议论文集中，第9939-9948页，*2021.2

用于三维场景的广义语义分割的神经语义域

补充材料

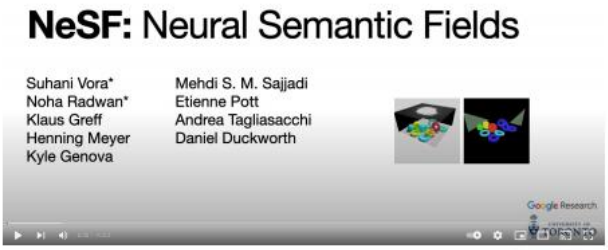

图6。概述视频-我们强烈建议在YouTube上观看高清概述视频。

A. 贡献

我们将在下面描述每个作者的具体贡献。

苏哈尼·沃拉是这项工作的两位第一作者之一。她实现了NeSF的2D推理框架；实现并运行了NeSF实验；实现并运行了DeepLab基线实验；建立项目网站；并撰写了论文的部分内容。

诺哈·拉德万是这部作品的两位第一作者之一。她实现了3D语义推理模块；实现并运行了NeSF实验；实现了NeSF的二维评估框架；建立项目网站；并编辑论文。

Klaus Greff负责Kubric [45]，在本工作中用于生成数据集的技术。他进一步协助开发数据集和概述视频中的动画。

亨宁·迈耶协助开发了ToyBox5和ToyBox13的数据集。他进一步为NeSF的代码库做出了重大贡献。

凯尔·热诺娃负责节约网络网络的基线。他进一步协助实验设计，并撰写了部分论文。

Mehdi S。M.Sajjadi提出了实验，并为NeSF建立的代码库做出了贡献

艾蒂安Pot实现了用于训练NeSF的可伸缩数据集管道。他进一步为NeSF的代码库做出了重大贡献。

安德里亚·塔利亚萨奇作为研究负责人领导了这个项目。他进一步提出了最初的NeSF模型架构建议实验；监督和撰写论文的大部分。

丹尼尔·达克沃斯作为一个技术负责人监督了这个项目。他协调了贡献者，设计并原型了实现NeSF的软件架构。他进一步提出、实现并运行了NeSF实验；生成本文使用的数据集；实现了NeSF的3D推理框架；写论文的部分；生成可视化；编写、组装和录制概述视频。

B. 培训详细信息

我们将在下面描述NeSF所使用的模型架构和训练过程及其基线。除非另有说明，我们使用所有的方法来训练每种方法

用每个场景中随机选择的9张图像来训练场景。对于KLEVR，这导致了100个火车场景；对于ToyBox5和玩具13的500个火车场景。所有的方法都是对从每个数据集的新场景中随机选择的4张图像进行评估的。对于KLEVR，这就产生了20个新场景；对于ToyBox5和ToyBox13，各有25个新场景。通过指定随机数生成器的种子，我们确保每种方法都能观察到每个场景中相同的随机选择的图像集。

NeRF。NeSF的第一阶段是对每个场景的NeRF模型的训练。我们采用了米尔登霍尔等人的模型架构和训练制度。[94].每个场景的密度场由一个包含8个256个单元的隐藏层的MLP描述，其外观由1个隐藏层和128个单元的额外MLP描述。我们使用10个八度音来进行位置编码。每个NeRF模型都是通过使用Adam优化器[67]从9个视图中随机选择的像素进行训练的。在25,000步的学习过程中，学习速率从1e-3呈指数级衰减到5.4e-4。我们在8个TPUv2核上训练每个NeRF模型大约20分钟。

NeSF。NeSF有两个主要的模型组件：一个3D UNet和一个MLP解码器。对于3D UNet，我们采用,,等人的UNet结构。[26]，删除了批规范层，只有2个最大池化操作。在每个最大池化之前，我们分别使用32、64和128个输出通道活动对于MLP解码器，我们使用了2个隐藏层，包含128个隐藏单元，每个单元都具有ReLU非线性。

我们用Adam优化器[67]训练NeSF。我们使用一个指数衰减的学习速率，初始化到1e-3，并在超过25000步中衰减到1e-5。在每一步，我们采用分层抽样的方法：我们随机选择32个

场景，然后从每个场景的训练摄像机中随机选择一组128个像素。对于体积绘制，我们根据NeRF [94]中使用的分层方法，沿着每条射线采样192个点。对于批处理中的每个场景，我们通过探测到64来离散NeRF的密度网格^{3}均匀间隔点。在离散化之前，我们会对每个场景应用一个围绕z轴（向上）的随机旋转。对于光滑正则化，我们均匀采样8192个addi-
来自每个场景的任意三维坐标，并添加标准差为0.05的随机噪声。当计算损失时，我们将权重分配为0。1到光滑度正则化项。

我们发现，我们能够在32个TPUv3核上训练NeSF在大约45分钟内收敛。

深度实验室。我们训练了一个DeepLab宽ResNet-38模型[138]，从COCO预先训练的检查点开始温暖。对于我们的优化方案，我们应用SGD +动量，缓慢的开始学习速率为1e-4，线性上升到6e-3，然后在55000步训练时，余弦计划衰减到1.26e-7。我们另外应用的重量衰减为1.0e-4。对于每个列车步骤，我们使用32的批处理大小。模型在32个TPUv3芯片上进行训练。为了能够重用性能良好的超参数配置，我们将输入图像从256x256上采样到1024x1024，对RGB输入使用双线性插值，对相应的语义映射使用最近邻插值。共享ConvNet。我们的稀疏ConvNet[43]实现是基于TF3D [58]和2D3DNet [41]实现的。除最后一层外，每个卷积层都是占用归一化的3x3x3稀疏空间卷积，然后是批处理范数，然后是ReLU。最后一层省略了批处理范数和ReLU。每个编码器阶段是一对卷积层，然后是一个2x2x2的空间最大池操作，每个解码器层是一个体素解池操作，然后是一对卷积层。编码器特征宽度分别为（64、64）、（64、96）、（96、128）、（128、160）、（160、192）、（192、224）、（224、256）。这些是每个块的第一和第二卷积层的输出通道计数。瓶颈是两个宽度为256的卷积层的序列。解码器特征宽度为（256、256）、（224、224）、（192、192）、（160、160）、（128、128）、（96、96）、（64、64）。最后，我们应用三个卷积层的大小(64,64，类_计数)，然后是一个软tmax层和一个交叉熵损失函数。我们的输入特性只是占用率（即，在所有输入点上的a1）。我们在一个[- 1,1]立方体场景中使用0.00个5宽的体素。我们使用SGD优化了45万步，动量为0.9，批量大小为5，初始学习率为0.0015，以及从步骤200,000开始到步骤450,000结束的余弦学习率衰减。我们添加一个！2重量衰减损失为1e-4，并在20 NVIDIA V100gpu上异步训练。我们应用以下数据增强： XY旋转高达10度，z旋转为180度，和a±±
随机尺度因子在0.9到1之间。1.

C.分析

.1.C定性结果

在图12、图13和图14中，我们对本文中研究的每个数据集进行了随机选择的定性结果。在每一行中，我们描述了地面真实RGB、深度和语义映射以及由NeSF、DeepLab和SparseConvNet生成的2维分割映射。我们观察到，所有的方法都能有效地分离前景物体从地板和背景。与稀疏通信网络不同，NeSF和DeepLab在正确的类别不明确时，倾向于将同一对象的不同部分分配给不同的语义类别。

虽然NeSF和SparseConvNet在设计上是多视图一致的，但对于像DeepLab这样的2D方法却不是这样。在图7中，我们演示了一个3D不一致的实例。在这个例子中，NeSF和SparseConvNet从两个视图中标记橙色沙发和白蓝色显示相同，而DeepLab的分类发生了变化。

NeSF三维密度场质量。值得注意的是，稀疏通信网络和NeSF之间的一个关键区别是提供了一个地面真实点云作为稀疏通信网络的输入。稀疏通信网络的几个方面可能有助于其相对于NeSF的整体优越性能，包括访问oracle 3D几何，稀疏点云输入表示，
或者稀疏网络网络模型架构。为了更好地理解存在的改进，我们首先通过视觉检查三维几何的三维密度场之间的差异与地面真实点云提供的场景选择KLEVR，ToyBox5和ToyBox13图8。我们观察到NeSF密度场经常忽略薄的结构和细节，而“飞蚊”在ToyBox13中尤为明显。通过改进NeRF表示来提高密度场的质量可以解决这类伪影。此外，类似于表7中的结果，这些改进可能会提高NeSF的性能。我们将三维输入表示和NeSF语义模型架构的替换留给未来的工作。

C..2消融术

模型消融术-表6。我们在ToyBox5模型上重复了我们的消融研究，并观察到与KLEVR模型消融的结果总体一致。表6显示了每个组件的结果。与KLEVR的结果相似，我们观察到以随机场景旋转的形式进行的数据增强提高的质量最大，分别增加了9.3%和6。分别为1%到2D和3D mIoU。探测的NeRF密度网格的空间分辨率再次被证实是至关重要的，并且明显在更大程度上大于KLEVR。我们假设这是在ToyBox5比KLEVR包含更精细的结构对象时发生的。

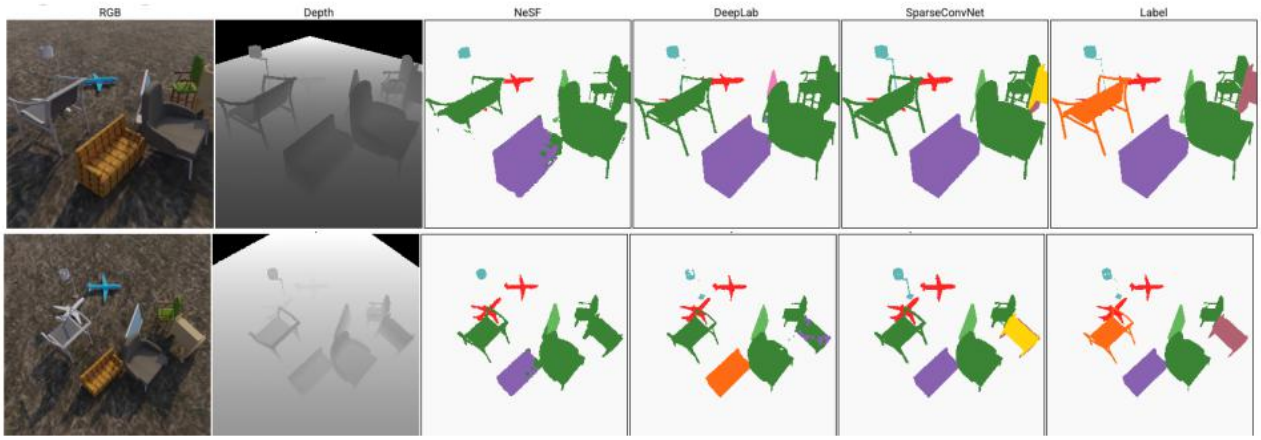

图7。多视图一致性虽然NeSF和SparseConvNet从同一场景的多个独立视图中对橙色沙发和显示器进行的分类相同，但DeepLab的预测各不相同。

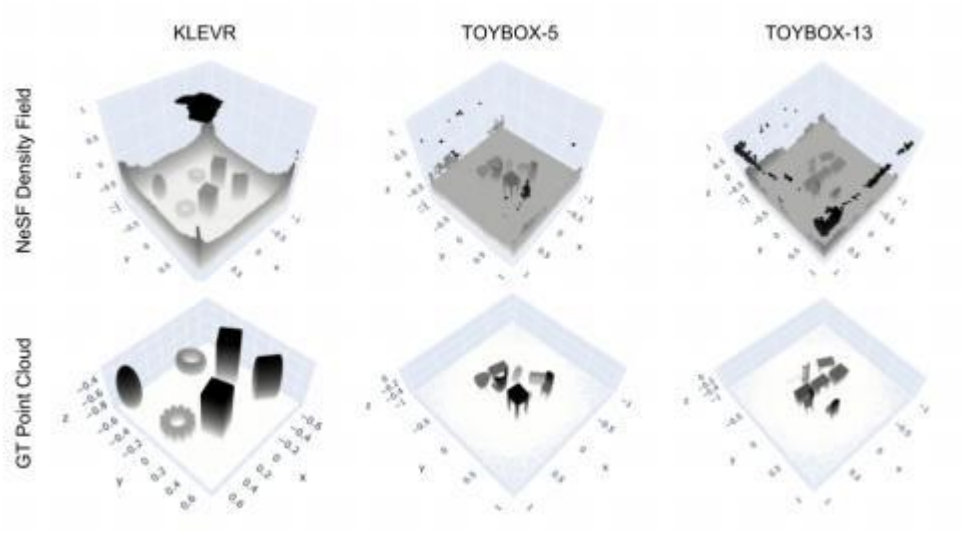

图8。KLEVR、ToyBox5和ToyBox13的NeRF三维密度场（顶部）和地面真实点云（底部）。我们发现NeRF的密度场准确地捕捉到了场景的三维几何形状。NeSF密度场以分辨率为128x128x128进行采样，并对3个数据集进行过滤，以便于可视化的正值，阈值分别为16、64和64。σ

对重建质量的敏感性-表7。我们重新评估了NeSF对NeRF重建的鲁棒性
在ToyBox5数据集的上下文中的质量。为了调制重建质量，我们将拟合NeRF模型时使用的RGB图像数量从5个变化到100个，并确认随着更多RGB图像的提供，NeRF重建质量有所提高。如前所述，NeSF的二维和三维分割质量随着NeRF的重建质量而单调提高。值得注意的是，当NeRF模型时，3D分割精度开始趋于接近88%
优化了50张RGB图像，性能在25到50张图像之间有很大的飞跃。

图9我们重复了我们的分析，提供了有限数量的语义标记NeSF对ToyBox5模型的训练地图。我们将所提供的标签地图的数量从1变化到50。与KLEVR设置类似，我们观察到为每个场景提供额外的语义映射可以提高性能，在5到10个映射之间会有很大的跳转。模型性能在每个场景大约25个地图时达到饱和。此外，该模型仍然能够概括每个场景只有1个语义映射。

.3.C多vew一致性

与传统的二维方法不同，NeSF的设计是三维一致的。在图10中，我们可视化了沿着NeSF语义预测的红色参考线绘制的外极平面。我们发现，最终的预测是

超参数		2D	3D
随机旋转	不	69.5	83.6
	是	78.8	89.7
密度网格	(32, 32, 32)	71.1	81.5
	(48, 48, 48)	76.4	89.3
	(64, 64, 64)	78.8	89.7
UNet	(16, 32, 64)	80.6	89.1
	(24, 48, 96)	80.1	89.8
	(32, 64, 128)	79.0	89.8
MLP	(0, 32)	78.6	90.7
	(1, 32)	79.7	89.8
	(1, 64)	80.7	89.4
	(2, 128)	79.2	89.5

表6。**消融：超参数-以随机场景旋转的形式增加了数据，提高了密度网格的空间分辨率，增加了UNet模型的容量，提高了二维和三维mIoU。**在来自ToyBox5数据集的500个场景上进行的实验。

# RGB图像	NeRF PSNR SSIM				NeSF 2D 3D	
5	±17:5	2:1	0:55	0:15±	15:0	17:9
10	19 ±2 :	2:9	0:62	0:15±	29:1	35:2
25	23 ±9 :	2:7	0:76	0:09±	61:4	74:1
50	26 ±3 :	2:1	0:81	0:06±	72:3	88:7
75	27 ±3 :	2:0	0:83	0:05±	72:6	89:5
100	27 ±9 :	2:0	0:84	0:04 73:6±		90:0

表7。**消融：ToyBox5对重建质量的灵敏度-我们的方法的准确性随着NeRF的重建质量而提高。**PSNR和SSIM在所有场景和聚合的指标中取平均值。对所有场景的实验
玩具盒5。

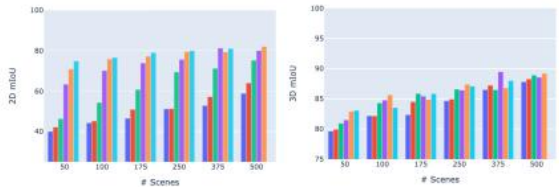

图9。消融：数据效率-2D和3D mIoU作为场景的列车场景数量的函数，每个场景监督1、2、5、10、25或50个语义映射。注意，50个语义映射不适合500个场景的内存，因此从结果中删除了这个特定的设置。NeSF概括到新场景，每个场景只有一个语义映射。每个场景的额外语义映射略微提高了准确性。在ToyBox5数据集上的实验。

一致的和光滑的，除非当一个浮子阻碍了相机的视野，如在EPI中的白色污点所示。在图11中，我们将看到这种现象的进一步示例。当摄像机围绕场景旋转时，一个浮动的密度质量阻碍摄像机的视图，由此产生的语义-

抽搐映射包含大量错误标记的像素。**我们强烈鼓励读者在相关视频中查看其他结果以获得更多细节。**

C..4限制

混淆矩阵在表8和表9中，我们在ToyBox13数据集上给出了NeSF的每类混淆矩阵，用于2D像素和三维点分类。而NeSF能够轻松识别更大的、清晰的语义类别，如橱柜、椅子、显示器或表(78.0-89.4% 2D，79。4-93.6% 3D)，它难以识别较小的物体类别，如步枪（56.3% 2D，75.3% 3D）或几何上未铰接的物体，如扬声器（38.5% 2D，40.4% 3D）。当NeSF混淆前景对象类别时，最常见的错误是在几何形状相似的类之间。例如，长椅经常被错误地贴上了椅子的标签。0% 2D，17.9% 3D)和沙发(26.5% 2D，27.4%
扬声器)，扬声器经常被错误地标记为表格（32.0% 2D，32.5% 3D）。

精度2D vs。3D.实验结果表明，NeSF在三维图像中的精度高于二维图像。我们发现这令人惊讶，特别是考虑到NeSF的语义监督的2D性质。我们认为最终的原因是由NeRF恢复的三维密度场中的“飞蚊症”。在表8中，我们看到每个语义类别中大约有510%的2D像素被错误地标记为“背景”。相比之下，表9显示，相同类型的误差在3D中出现的时间约为1%。最突出的例外是台类别，其对象通常包含NeSF难以捕获的薄结构。

浮子的影响。在图11中，我们定性地展示了“飞蚊器”如何降低NeSF在图像空间中的准确性。在这一组5个视频帧中，我们演示了一个摄像机路径通过　　在一片漂浮的云前。该云被分配给背景语义类别，并模糊了场景中的前景对象。尽管这些对象是正确的　　被标记后，生成的语义映射是不正确的。因此，NeSF获得的2D mIoU低于3D mIoU，因为后者不受飞蚊的阻碍，并由表8和表9证实。我们相信，在几何构造中消除这种故障将显著提高NeSF的精度。解决方案是很容易提供的方法构建-　在NeRF [141]上。**我们强烈鼓励读者**
查看附带视频中的附加结果以获得更多细节。

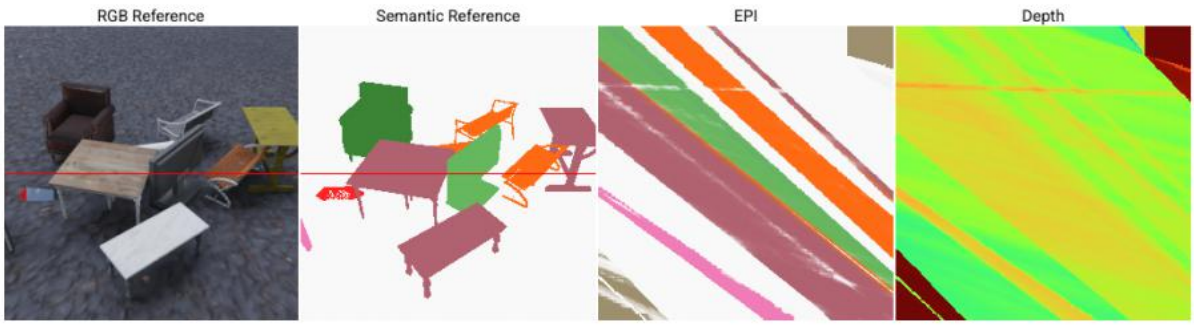

图10。我们的外极平面通过在相机从右向左移动时沿红色扫描线渲染外极平面来证明NeSF的三维一致性。外极平面是平滑和一致的，除非“漂浮物”通过相机前面。

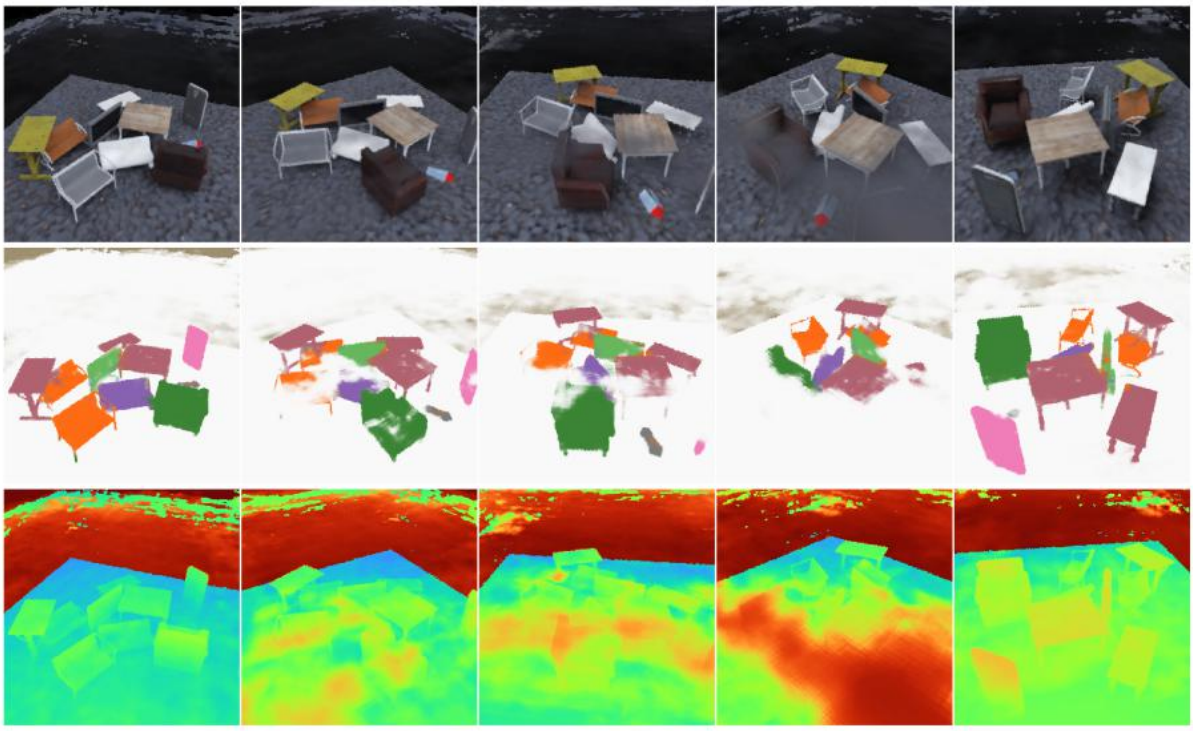

图11。在上面的3行中，我们举例说明了NeRF的RGB重建（顶部），NeSF的语义场，和NeRF的密度场（底部）。当NeRF的密度字段包含“飞蚊器”时，NeSF经常将它们分配给背景语义类别。

	背景	飞机	长凳	内阁	小	椅子	陈列	灯	扩音器	步枪	沙发	表	电话	容器
背景	99.1%	0.1%	0.0%	0.0%	0.0%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%
飞机	12.9%	69.2%	3.2%	0.1%	0.0%	1.9%	0.0%	0.3%	0.0%	1.4%	0.0%	4.6%	0.0%	6.3%
长凳	9.1%	0.7%	42.1%	0.1%	0.0%	17.0%	0.0%	0.2%	0.0%	1.2%	26.5%	1.7%	0.0%	1.4%
内阁	2.2%	0.0%	0.0%	78.0%	0.0%	0.1%	2.7%	0.0%	12.8%	0.0%	0.0%	2.7%	1.5%	0.0%
小汽车	5.0%	0.1%	0.1%	0.0%	84.6%	0.5%	0.1%	0.3%	0.3%	0.2%	1.3%	4.1%	0.0%	3.3%
椅子	5.8%	0.0%	0.9%	0.0%	0.0%	89.4%	0.1%	0.0%	0.1%	0.0%	2.9%	0.7%	0.0%	0.0%
陈列	9.2%	0.0%	0.3%	0.4%	0.1%	1.2%	83.3%	0.0%	3.7%	0.0%	0.2%	1.3%	0.2%	0.0%
灯	11.1%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%	61.4%	18.1%	0.0%	0.0%	0.1%	7.8%	0.8%
loudspeake	5.4%	0.1%	4.3%	10.5%	0.0%	0.6%	3.2%	1.1%	38.5%	0.0%	2.2%	32.0%	1.7%	0.4%
步枪	30.7%	1.6%	2.0%	0.2%	0.0%	0.7%	3.1%	0.0%	0.0%	56.3%	3.0%	0.7%	0.0%	1.7%
沙发	10.0%	2.3%	0.6%	0.0%	0.0%	21.6%	0.0%	0.1%	0.1%	0.0%	61.6%	0.0%	0.0%	3.6%
表	5.3%	0.0%	1.7%	3.1%	0.0%	4.6%	0.1%	3.1%	1.2%	0.0%	0.7%	79.9%	0.0%	0.3%
电话	1.7%	0.0%	0.0%	9.3%	0.0%	0.2%	3.7%	0.0%	15.7%	0.0%	0.0%	0.3%	69.0%	0.0%
容器									0.0%					

表8。利用ToyBox13上的NeSF进行二维语义分割的混淆矩阵。每一行对应一个地面真实标签，并被归一化为之和为100%。NeSF最常见的错误包括混淆形状相似的物体，以及将小物体和薄物体分类为背景。正确的分类将以粗体突出显示。

	背景	飞机	长凳	内阁	小汽车	椅子	陈列	灯	扩音器	步枪	沙发	表	电话	容器
背景	99.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%
飞机	1.7%	78.0%	3.6%	0.0%	0.0%	1.8%	0.0%	0.2%	0.0%	2.1%	0.0%	4.8%	0.1%	7.6%
长凳	6.0%	0.8%	42.7%	0.2%	0.0%	17.9%	0.0%	0.3%	0.0%	1.7%	27.4%	1.6%	0.0%	1.4%
内阁	0.4%	0.0%	0.0%	79.4%	0.0%	0.0%	2.9%	0.0%	13.0%	0.0%	0.0%	2.6%	1.7%	0.0%
小汽车	0.8%	0.4%	0.3%	0.0%	86.7%	0.6%	0.0%	0.2%	0.4%	0.3%	2.1%	4.0%	0.0%	4.3%
椅子	1.0%	0.0%	1.1%	0.0%	0.0%	93.6%	0.1%	0.1%	0.1%	0.1%	3.1%	0.8%	0.0%	0.0%
陈列	0.6%	0.1%	0.4%	0.3%	0.0%	1.0%	90.7%	0.0%	4.8%	0.0%	0.4%	1.4%	0.3%	0.0%
灯	2.3%	0.0%	0.0%	0.1%	0.1%	0.4%	0.0%	66.7%	20.7%	0.0%	0.0%	0.1%	8.8%	0.9%
loudspeake	1.1%	0.1%	4.4%	12.5%	0.1%	0.4%	3.0%	1.2%	40.4%	0.0%	2.1%	32.5%	1.9%	0.4%
步枪	2.4%	4.9%	4.6%	0.6%	0.1%	1.8%	3.6%	0.0%	0.0%	75.3%	3.2%	0.9%	0.0%	2.7%
沙发	0.7%	2.9%	0.8%	0.0%	0.0%	22.8%	0.0%	0.1%	0.1%	0.0%	68.2%	0.0%	0.0%	4.3%
表	0.6%	0.0%	1.8%	3.5%	0.0%	4.7%	0.0%	3.2%	1.5%	0.1%	0.8%	82.8%	0.0%	0.8%
电话	0.1%	0.0%	0.0%	10.5%	0.0%	0.0%	4.5%	0.0%	15.8%	0.0%	0.0%	0.0%	69.1%	0.0%
容器	1.1%	13.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.4%	0.6%	0.0%	0.0%	80.7%

表9。利用ToyBox13上的NeSF进行三维语义分割的混淆矩阵。每一行对应一个地面真实标签，并被归一化为之和为100%。NeSF最常见的错误包括混淆形状相似的物体，以及将小物体和薄物体分类为背景。正确的分类将以粗体突出显示。

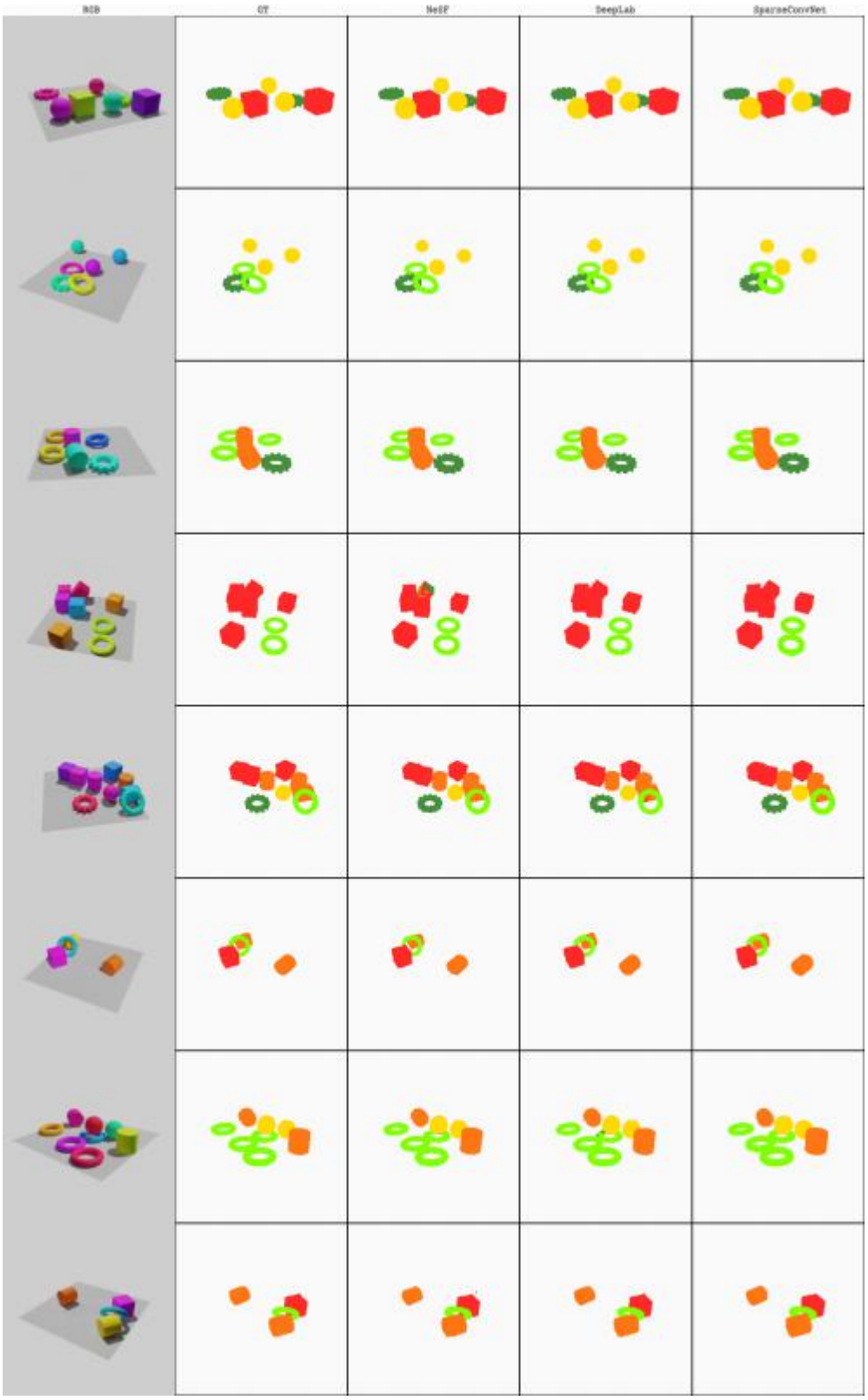

图12。关于KLEVR的附加定性结果

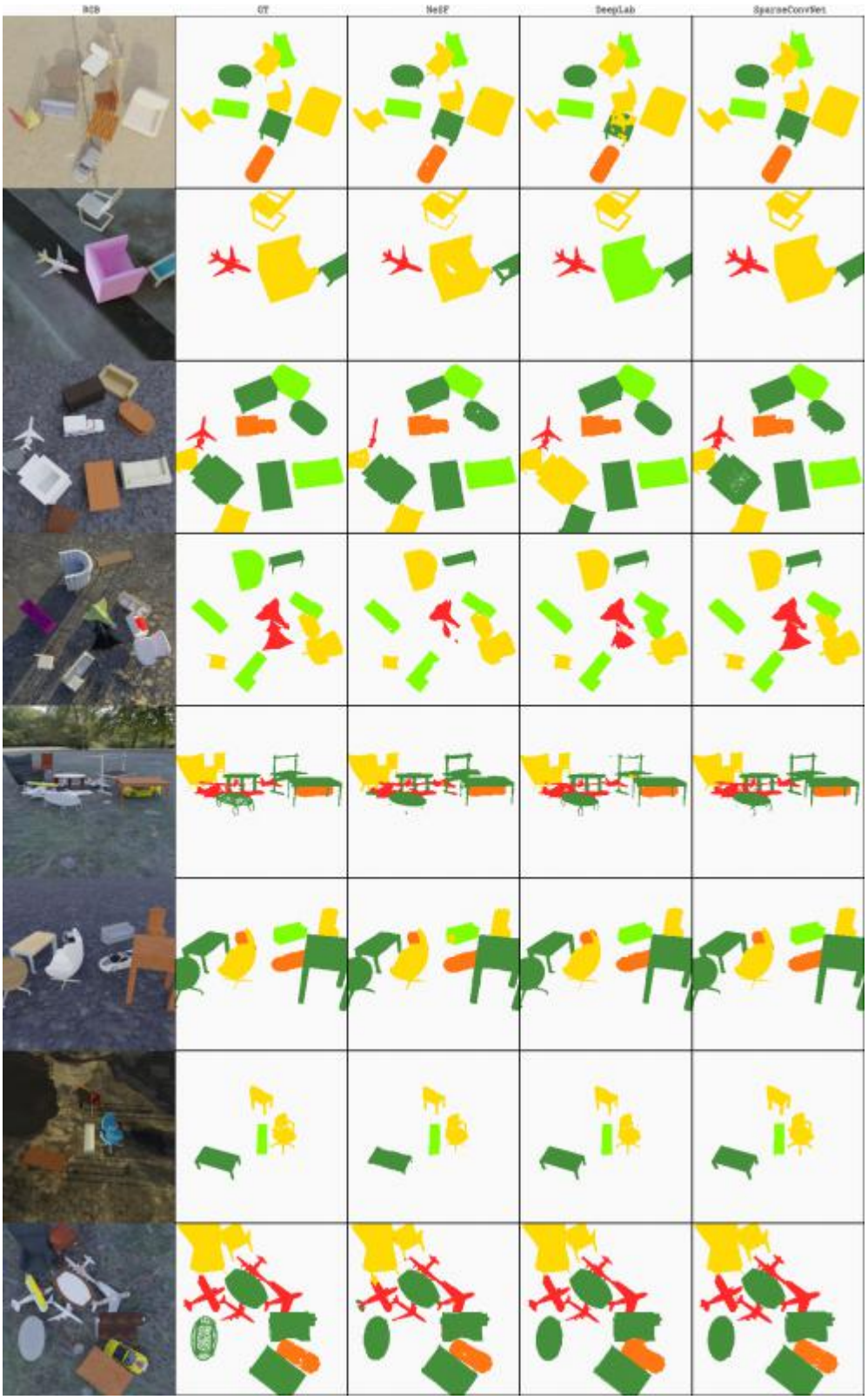

图13。关于玩具盒5的附加定性结果

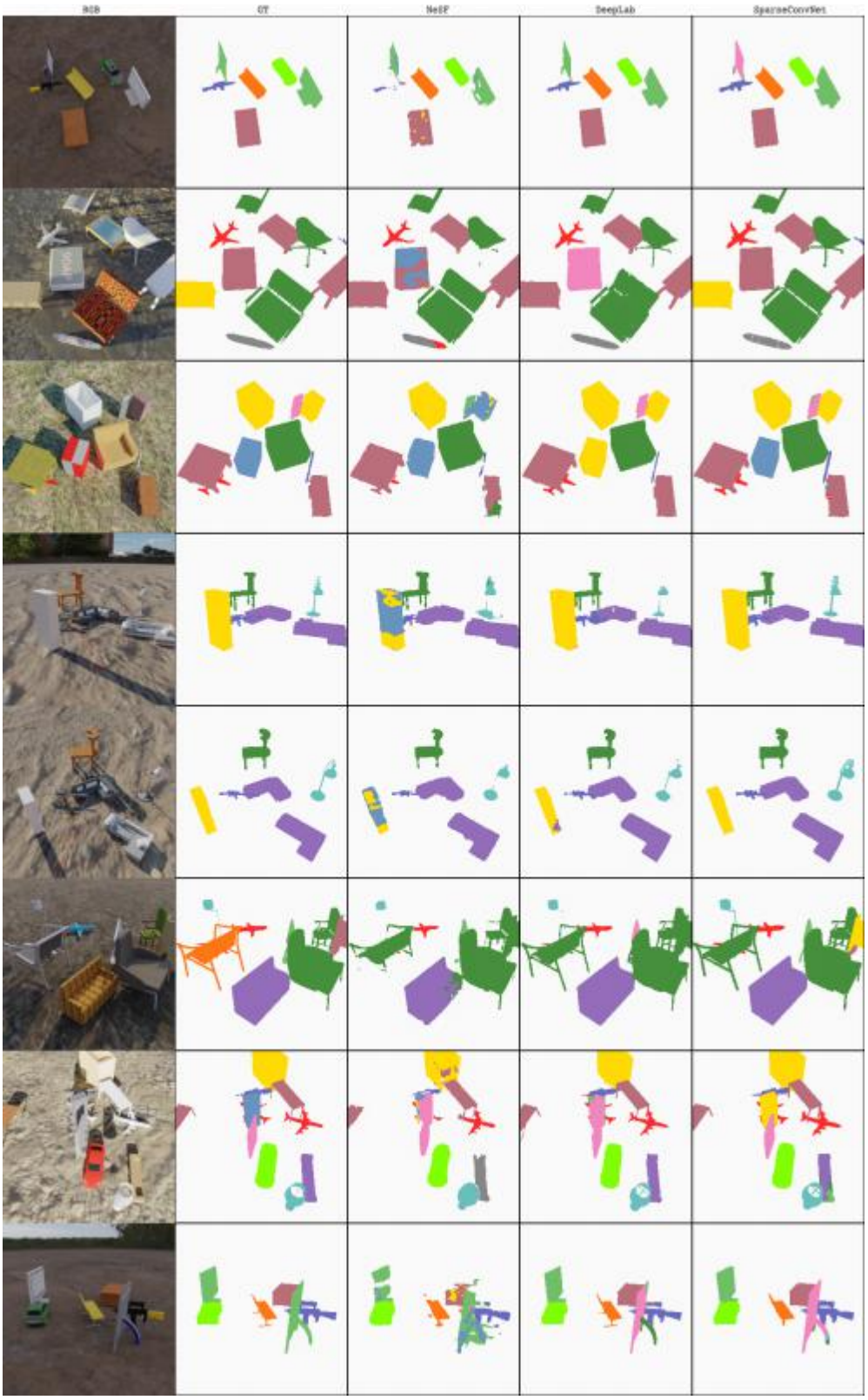

图14。关于玩具盒13的额外定性结果