GARField

题目：GARField: Group Anything with Radiance Fields
来源：UC Berkeley 和 Luma AI
项目： https://www.garfield.studio/

文章目录

- 摘要
- 一、前言
- 二、相关工作
- 2.1 层次分组
    - 2.2 NeRF的分割
    - 2.3 3D 特征场
- 三、method
- 3.1 2D Mask 生成
    - 3.2 Scale-Conditioned Affinity Field(尺度条件亲和场)
    - 3.2.1 对比监督
        - 3.2.2 密集尺度监督
        - 3.2.3 射线和掩码采样
    - 3.3 实施细节
- 四、层次分解
- 五、实验
- 5.1 场景分解（定性）
    - 5.2 层次结构（定量）
- 六、局限性
- 七、代码
- 总结

摘要

提示：这里可以添加本文要记录的大概内容：

分组(或者分割)本身是模糊的，因为在不同粒度级别上，场景的分割标准不同——**挖掘机的车轮应该被认为是独立的还是整体的一部分**？本文提出 **辐射场分组 GARField，一种将三维场景，从带pose图像的输入分解为语义组的方法** 。方法通过物理尺度来接受群体的模糊性：通过优化一个按尺度划分的3D密切特征场，从SAM模型提供的二维mask来优化，以从粗到细的层次结构，通过自动树构造或用户交互推导出可能分组的层次结构。

GARField能够实现 **对象的集体**、**对象**和**各种子部分**， 具有令人兴奋的下游应用程序，如3D资产提取或动态场景理解。

一、前言

如图1，虽然NeRFs 等技术可以恢复场景的逼真的3D重建，但世界被建模为一个没有结构意义的单一体积。作为人类，我们不仅可以重建场景，但我们也有能力组在多个层次的粒度，分类理解场景。

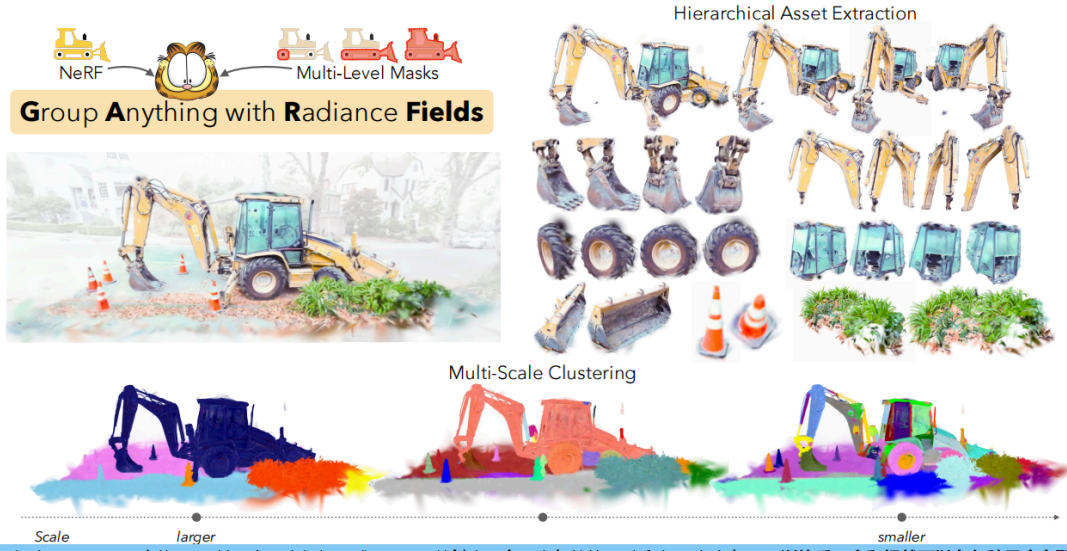

图1：提出GARField，它将mask基于多层次组提取成NeRF，以创建一个尺度条件的3D亲和场（左上角）。**训练后，亲和场就可以在各种尺度上聚类，以不同的粒度级别分解场景**，比如将挖掘机分解成它的子部分（底部）。三维资产可以通过**场景自动提取**或通过**用户点击**从层次结构中提取，如这里所示（右上角）

提出了GARField方法，给定姿态的图像，**重建一个三维场景和一个 scale-conditioned affinity field，使将场景分解成组的层次结构。**例如，GARField可以提取整个挖掘机（图1右上）以及它的子部件（右下）。 这种密集的层次3D分组使诸如3D资产提取和交互式分割等应用程序成为可能。

GARField将一组二维分割mask，提取成一个三维体积尺度条件的亲和场。因为分组是一项模糊的任务，二维标签可能是重叠或冲突的，导致了挑战。 我们通过**利用一个具有尺度条件的特征域**来解决问题。具体地说，GARField优化了一个密集的三维特征场，它被监督，使特征距离反映了点的亲和力。尺度调节使两点在大尺度上具有较高的亲和力，而在较小尺度上具有较低的亲和力，如图2所示 。

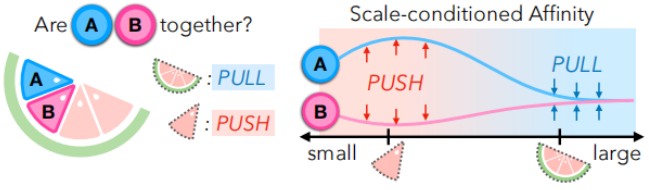

图2.**分组时尺度的重要性**。单个点可能属于多个组。GARField使用尺度条件反射，将这些冲突的信号调和成一个亲和场。

我们用SAM得到输入图像的候选分割mask。对于每个mask，基于场景几何计算一个物理比例，利用3D尺度来解决视图或候选掩模之间的不一致。

一个行为良好的亲和场de特点：

> 1)**可和性**，这意味着如果两个点与第三个点相互分组，它们本身应该分组在一起；2)**包容**，这意味着如果两个点在一个小的尺度上分组，它们应该在更高的尺度上分组在一起。

GARField使用的**对比损失**和**抑制辅助损失**鼓励了这两种特性。

二、相关工作

2.1 层次分组

从前景分割开始，二维图像的研究一直很广泛。有几种方法基于光谱聚类的思想，通过经典的纹理线索来提取轮廓，并通过一个自顶向下的[37]或自下而上的模型，用于多层次分割和更复杂的层次场景解析[1,25,31]。

许多工作通过定义一组类别来规避分组中的模糊性问题，其中的实例将被分割，即全景分割[10,14]。最近，SAM 将这种模糊性off-loads到提示中，每个像素上可以提出多个分割掩模。然而，SAM不能在场景中恢复一组一致的层次组（我们通过多尺度三维蒸馏实现）

我们的方法从2维模型中提取信息：考虑完整的场景，并专注于3D对象。

2.2 NeRF的分割

现有的NeRF中的分割方法通常通过使用地面真值语义标签[29,38]，匹配实例掩码[18]，或在NeRF[34]上训练三维分割网络，将分割掩模提炼成三维分割网络。但是，这些技术不考虑层次结构分组，而只对对象或实例的平面层次结构感兴趣。Ren等人[27]利用图像涂鸦的形式的人类互动来分割对象。最近，Cen等人[3]试图通过用户提示跟踪相邻视图之间的2D掩模，从SAM中恢复3D一致掩模。Chen等人[4]尝试通过将SAM编码器特征提炼成3D并查询解码器。与这些方法相比，**GARField不需要用户输入；它能够自动获得场景的分层分组，而且恢复的组根据定义是视图一致的。**

2.3 3D 特征场

将高维特征分解成一个神经场，与辐射场（视角相关的颜色和密度）相结合，已经被彻底探索。Semantic NeRF [38]、蒸馏特征场[16]、神经特征融合场[33]、Panoptic Lifting[29]等方法，将三维特征场优化的逐像素二维特征提炼成三维，重建体积渲染后的二维特征。这些特征可以来自预先训练好的视觉模型，如DINO或来自语义分割模型。LERF [13]将这一想法扩展到一个有尺度条件的特征领域，使其能够从像CLIP这样的全局图像嵌入中训练特征域。

GARField同样在三维空间，优化了尺度条件特征字段；然而，多尺度特征的目的是解决分组中的歧义，而不是像CLIP那样重建显式的二维特征。此外，LERF没有空间分组。上述方法都是基于对图像特征的直接监督，而其他方法，如NeRF-SOS [8]和对比Lift [2]，使用基于相似性的射线对之间的对比损失，在单一尺度上优化任意特征场。GARField使用这种对比的方法，因为它允许基于掩码标签定义点之间的成对关系。然而，我们设计了一个尺度条件下的对比损失，它允许提取相互冲突的mask 到 3D。

三、method

3.1 2D Mask 生成

GARField以pose图像为输入，生成一个分层的三维场景分组，以及一个标准的三维体积辐射场和一个有尺度条件的亲和场。首先用SAM得到输入图像的候选mask。接下来，**通过输入的三维位置和欧式尺度，优化一个体积辐射场和亲和场，并输出一个特征向量。**亲和度是通过比较点对的特征向量来获得的。优化后，生成的亲和字段可以用于分解场景，通过以粗到细的方式递归地聚类三维特征嵌入，或者

用于分割用户指定的查询。整个管道如图3所示。

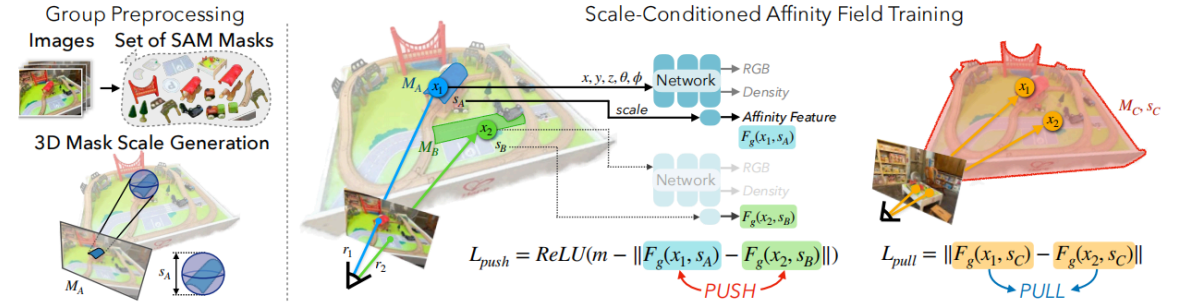

图3。GARField方法：（左）给定输入images，密集查询SAM来提取一组候选mask，并通过从NeRF中逆投影深度，来**为每个候选mask分配一个物理尺度Sa**，被用来训练某个尺度条件下的亲和场（右）。训练时，如果成对采样的光线位于不同的mask，它们会被推开；反之则会被拉在一起。亲和力只在每个mask的尺度上被监督，这有助于解决冲突。

2D mask筛选：首先用SAM的自动掩码生成器，得到图像的二维mask候选对象，为每个mask分配一个3D尺度。具体的，在一个点网格中查询SAM，并在每个查询点产生3个候选分割掩码。然后，**通过置信度过滤这些掩模，并删除几乎相同的掩模**，以产生多个大小的候选掩模列表，可以重叠或包括彼此。这个过程是独立于视点完成的，产生的mask可能不一致。目标是生成一个基于对象的物理大小的分组层次结构。因此，**我们为每个2D mask分配了一个物理三维尺度，如图3所示。**为此，我们部分地训练了一个辐射场，并渲染了一个来自每个训练摄像机pose的深度图像。接下来，**对于每个mask，我们考虑该掩模内的三维点，并根据这些点的位置分布的范围来选择比例。**该方法保证了掩模的三维尺度存在在相同的世界空间中，实现了尺度条件下的亲和力。

3.2 Scale-Conditioned Affinity Field(尺度条件亲和场)

尺度条件是GARField的一个关键组成部分，它允许整合不一致的二维掩码候选：相同的点可能取决于所需分组的粒度。尺度条件减轻了这种不一致性，因为它解决了查询应该属于哪个组的歧义。在尺度分割条件下，同一点的冲突掩模在训练过程中不再相互对抗，而是在不同的亲和尺度下在同一场景中共存。

我们在三维点 **x** 和欧氏尺度 **s** 上定义了尺度条件亲和场 $\mathbf{F_g(x,s)\rightarrow R^d}$，类似于LERF [13]。输出特性被限制在一个单位超球体内，在一个尺度上两点之间的亲和性由 $\mathbf{A(x_1,x_2,s) = -||F_g(x_1,s)-F_g(x_2,s)||_2}$ 定义。这些特征可以使用基于NeRF密度的相同渲染权重，以加权平均值进行体渲染，以获得每条射线的值。

3.2.1 对比监督

根据DrLIM [9]，采用**margin-based contrastive 对比目标**进行监督。损失又两部分组成：给定的尺度下，同一组中的特征接近，不同组中的特征分开。

具体来说，从同一训练图像中采样掩模 $\mathbf{M_A}$、$\mathbf{M_B}$ 的两条射线$\mathbf{r_A}$、$\mathbf{r_B}$，以及相应的尺度 $\mathbf{s_A}$和$\mathbf{s_B}$。我们可以沿每条射线，以体渲染方式得到尺度条件的亲和特征 $\mathbf{F_A}$和$\mathbf{F_B}$。如果$\mathbf{M_A=M_B}$，特性将通过L2距离拉在一起：$\mathbf{L_{pull}= ||F_A-F_B||}$；反之，特性将被分开：$\mathbf{L_{push} = ReLU(m-||F_A-F_B||)}$，其中m是下界距离或边界。这种损失只适用于从同一图像中采样的射线，因为在不同视点上的掩模没有对应关系。

3.2.2 密集尺度监督

仅有对比损失的监督，并不足以维持尺度分层。我们引入以下修改来解决：

持续尺度监督。使用3D mask尺度，分组只在mask对应的离散像素点处定义。这导致了大的无监督区域，如图9顶部所示。我们**通过在当前的mask尺度和第二个最小的mask尺度之间均匀随机地扩大尺度s来加强尺度监督。**当射线mask是给定视点的最小mask时，我们在0和s_0之间进行插值。确保了在整个领域的持续规模监督，没有留下无监督的区域

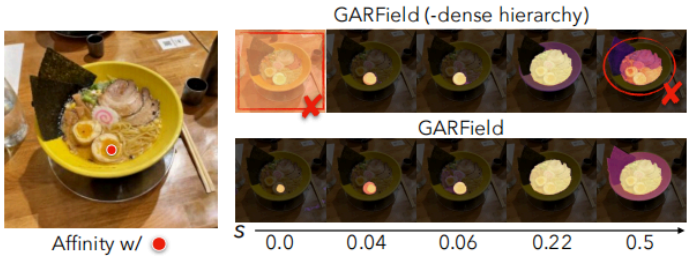

图9：消融实验：没有密集的层次监督，点可能在不同尺度上有不一致的亲和关系：
1)在无监督尺度上存在虚假的大亲和度，
2)在较大尺度上亲和力的意外下降。

遏制辅助损失：如果两条射线 r_1 和 r_2 在同一个尺度为s的掩模中，那么它们也应该在任何大于s的尺度上被拉在一起。每个训练步骤中，对于以s尺度分组的光线，我们另外采样一个更大尺度的 s'>s，光线也被拉在一起。这确保了在小尺度上的亲和性在大尺度上不会失去。

3.2.3 射线和掩码采样

为平衡图像的数量和用于监督的点对的数量，每次采样16张图像，每幅图像采样256个点，每次序列迭代得到4096个样本

对于每个采样的射线，还必须选择一个mask作为训练的组标签。在每个训练步骤中，我们从每个光线对应的mask列表中随机选择一个mask。

> 1)选择mask的概率与掩模的二维像素面积的对数成反比，**防止大尺度控制采样过程**，因为可以通过更多的像素来选择更大的掩模。

> 2)在mask选择过程中，我们协调同一图像中光线选择的随机尺度，以增加正对的概率。为此，我们对每幅图像采样一个介于0到1之间的单个值，并以相同的值索引到每个像素的掩码概率CDF中，以确保位于同一组内的像素被分配相同的掩码。

3.3 实施细节

该方法是在Nerfacto[32]的基础上，为 grouping field 定义一个单独的输出头。 grouping field 用24层的hashgrid[23]表示，每层特征维为2，有256个神经元和ReLU激活的4层MLP表示，以scale作为额外输入。我们将相机的范围限制在2×，并使用sklearn的 quantile transform对三维mask尺度分布的MLP输入进行归一化（第3.1节）。输出嵌入件的维数为d = 256维。来自亲和特性的梯度不影响来自NeRF的RGB输出，因为这些表示不共享任何权重或梯度。

经过2000步NeRF优化后，开始训练 grouping field，给出几何时间收敛。为了加速训练，**首先体渲染哈希值，然后将其作为MLP的输入，以获得射线特征**。使用这种延迟渲染，可以只用一个额外的MLP调用，以不同的尺度查询相同的射线。在输入MLP之前，我们将体积呈现的结果归一化为单位范数，对于点级查询，单个哈希网格值被归一化。预处理SAM大约需要3-10分钟，然后在GTX 4090上进行大约20分钟的训练

四、层次分解

一旦优化了尺度条件的亲和力，**GARField生成一个3D groups的层次结构，组织在树中，这样每个节点就被分解成潜在的子组。**为了做到这一点，我们通过减少亲和力的尺度来递归地聚类组，使用HDBSCAN [19]，这是一种基于密度的聚类算法，不需要先验集群的数量。

这种聚类过程，可以在二维中对生成mask的图像中的体渲染特征进行，**或者在三维跨点中生成点云。**

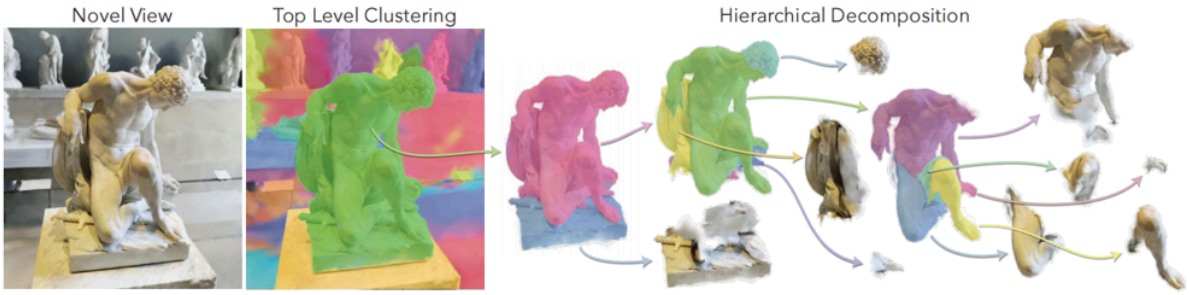

图6. 3D分解： GARField可以递归查询，将场景聚为对象及其子部分。

Initialization hierarchy：首先以一个大规模尺度 s_{max} 全局集群特征（实验设置为1.0，对应于输入摄像机的位置的程度），作为场景分解中的top节点。

递归聚类：为了生成场景节点的层次树，我们**迭代地减少一个固定的epsilon（我们使用0.05），在每个叶节点上运行HDBSCAN。**如果HDBSCAN为给定节点返回多个集群，那么我们将这些集群添加为子集群并递归。**持续到尺度0**，此时过程终止，返回当前树。

五、实验

现有的三维扫描数据集，倾向于关注对象级扫描，是模拟的，或主要包含室内家庭场景[6]。**为了评估GARField，我们使用了来自 Nerfstudio 和 LERF数据集的各种室内和室外场景。**图3和图6提供了定性结果。

5.1 场景分解（定性）

使用Gaussian Splatting [12]通过查询高斯中心的GARField亲和场来可视化分解。我们这样做是因为与nerf相比，Gaussian Splatting在3D中更容易分割。所有的渲染都是完整的3D模型，而不是2D图像视图的分割。

我们可视化了两种类型的分层聚类结果。**图7以手工选择的粗尺度对场景进行全局聚类，然后从聚类中，选择对应于少数对象的组，并将它们进一步分解为子组。**我们可视化了在连续递减的尺度上获得的簇，这增加了组的粒度。GARField实现高保真3d分组在广泛的场景和对象，从人造对象，如键盘，复杂的自然对象像植物，可以分组个体花以及他们的花瓣和叶子。通过改变尺度，在不同的层次上分离物体，例如花盆中的每一片叶子（左第一行）。

图7.**使用 Gaussian Splats产生三维可视化**：从GARField中，我们通过选择顶级集群从全局场景中提取对象，然后以递减的尺度可视化其局部集群。GARField可以生成完整的3D mask，并根据输入mask将这些对象分解成有意义的子部分

5.2 层次结构（定量）

使用两个指标进行定量评估：第一种测量来自多个视图中的**标签的视图一致性**，第二次通过mIOU对地面真实人类注释，**测量各种层次mask的召回。**

三维完整性：对于下游任务，组对应于完整的三维对象是很有用的，例如，包含整个对象而不是它的某一侧的组。虽然GARField总是通过构造生成与视图一致的组，但它可能不一定包含完整的对象。我们通过检查整个3D对象是否跨一系列视点组合在一起来评估其完整性。为了做到这一点，在5个场景中，选择一个3D点投影到3个不同的视点，并标记3个相应的视图一致的真实mask，包含在粗、中和精

细水平的点。在这些点上，我们以0.05的增量从GARField的多个尺度上挖掘多个掩模，在每个尺度上，基于0.9的特征相似性阈值获得一个mask。我们还通过点击图像中的点并拍摄所有3个面具来与SAM进行比较。我们报告了两种方法对所有候选mask计算的最大mIOU

结果如表1所示。GARField在跨视点上比SAM生成更完整的3D掩码，从而产生具有多视图人工对象注释的更高的mIOU。这种效果在最细粒度的层面上尤其明显。

Scene	Fine		Medium		Coarse	
	SAM	Ours	SAM	Ours	SAM	Ours
teatime	81.6	**92.7**	97.3	**97.9**	-	-
bouquet	17.4	**76.0**	73.5	**81.6**	76.1	**85.4**
keyboard	65.3	**88.8**	73.6	**98.4**	-	-
ramen	53.3	**79.2**	74.7	**90.7**	92.6	**95.5**
living_room	85.3	**90.5**	74.2	**80.7**	88.6	**94.4**

Table 1. **3D Completeness.**

Scene	SAM [15]	Ours (-scale)	Ours (-dense)	Ours
ramen	74.9	64.1	74.1	**85.6**
teatime	64.9	67.7	66.1	**86.6**
keyboard	23.2	57.6	73.1	**77.9**
bouquet	34.4	49.8	72.9	**76.4**
living_room	59.6	49.7	62.1	**76.6**

Table 2. **Hierarchical Grouping Recall**

层次分组召回：测量GARField在多个粒度上的召回。在5个场景中，我们选择一个新的视点，并为1-2个对象标记多达3个Grountruth 层次组。GARField通过聚类图像空间特征，输出一组如第4节所述的mask，每个树节点输出一个mask。我们通过保留所有输出mask与SAM的自动mask生成进行了比较。我们通过两种方式删除了GARField：GARField（-尺度）删除了尺度划分；而GARField（-层次结构）删除了密集监督

消融表明，规模条件反射和规模致密化对于高质量的分组是必要的。图9显示了在单纯监督下的更高规模的亲和分解。

六、局限性

GARField的核心是从2D mask生成器中提取输出，因此**如果2D mask不能包含所需的组，这将不会出现在3D中。视点不均匀的区域可能会出现人为的群体边界**，例如，如果一个对象只从近距离观察，它可能永远不会被分组在一起，因为没有输入视图完整地包含它。我们使用物理大小来处理组的模糊性，但在一个尺度内可能有多个分组。例如，与容器中包含的对象可能会发生冲突，因为有该对象和没有该对象的容器可以具有相同的比例。未来的工作可以考虑其他方法来解决分组歧义，如功能支持。比例条件反射的另一个结果是，不同大小的对象部分分别从树中分支，而不是一次分支：同一个表上的多个对象可能出现在树的不同层次上。本工作中的树生成是一种单纯的贪婪算法，它可以导致更深层次的虚假的小群，如补充部分中的树所示。未来的工作可能会探索更复杂的层次集群的方法。