GARField: Group Anything with Radiance Fields

Chung Min Kim*[1] Mingxuan Wu*[1] Justin Kerr*[1] Ken Goldberg[1]
Matthew Tancik[2] Angjoo Kanazawa[1]
∗ Denotes equal contribution
[1]UC Berkeley [2] Luma AI

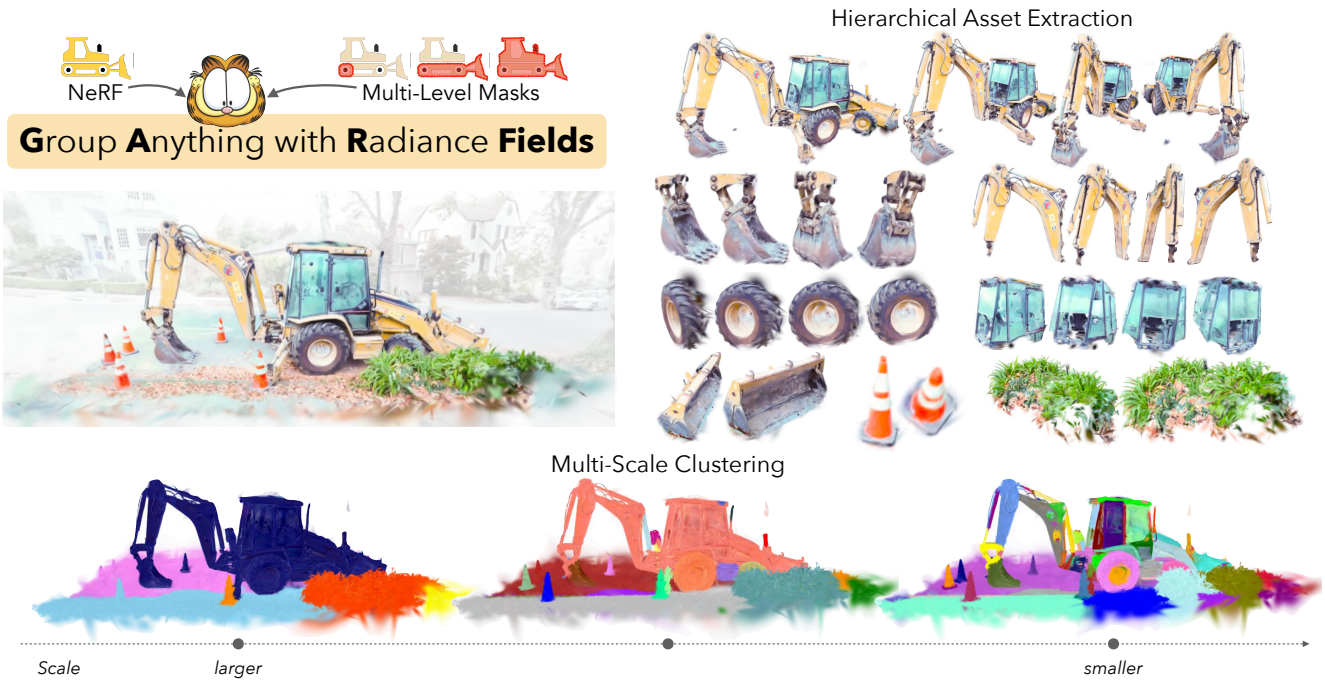

Figure 1. **Group Anything with Radiance Fields (GARField)**: GARField distills multi-level groups represented as masks into a NeRF to create a scale-conditioned 3D affinity field (top left). Once trained, this affinity field can be clustered at a variety of scales to decompose the scene at different levels of granularity, like breaking apart the excavator into its subparts (bottom). 3D assets can be extracted from this hierarchy by extracting every group in the scene automatically or via user clicks, as visualized here (top right).

Abstract

Grouping is inherently ambiguous due to the multiple levels of granularity in which one can decompose a scene—should the wheels of an excavator be considered separate or part of the whole? We propose Group Anything with Radiance Fields (GARField), an approach for decomposing 3D scenes into a hierarchy of semantically meaningful groups from posed image inputs. To do this we embrace group ambiguity through physical scale: by optimizing a scale-conditioned 3D affinity feature field, a point in the world can belong to different groups of different sizes. We optimize this field from a set of 2D masks provided by Segment Anything (SAM) in a way that respects coarse-to-fine hierarchy, using scale to consistently fuse conflicting masks from different viewpoints. From this field we can derive a hierarchy of possible groupings via automatic tree construction or user interaction. We evaluate GARField on a variety of in-the-wild scenes and find it effectively extracts groups at many levels: clusters of objects, objects, and various subparts. GARField inherently represents multi-view consistent groupings and produces higher fidelity groups than the input SAM masks. GARField's hierarchical grouping could have exciting downstream applications such as 3D asset extraction or dynamic scene understanding. Project site: https://www.garfield.studio/

1. Introduction

Consider the scene in Figure 1. Though recent technologies like NeRFs [20] can recover photorealistic 3D reconstruc-

tions of this scene, the world is modeled as a single volume with no structural meaning. As humans, not only can we reconstruct the scene, but we also have the ability to *group* it at multiple levels of granularity — at the highest level, we see the parts of the scene *i.e*. the excavator, bushes, and the sidewalk, but we are also able to decompose the excavator into its parts such as its wheels, crane, and the cabin. This ability to perceive the scene at multiple levels of groupings is a key component of our 3D understanding, enabling us to interact with the world by understanding what belongs together. However, these different levels of granularity introduce ambiguity in groups, making it a challenge to represent them in a coherent 3D representation. While there are multiple ways to break this ambiguity, we focus on the physical scale of entities as a cue to consolidate groups into a *hierarchy*.

In this work we introduce Group Anything with Radiance Fields (GARField), an approach that, given posed images, reconstructs a 3D scene along with a scale-conditioned affinity field that enables decomposing the scene into a hierarchy of groups. For example, GARField can extract both the entire excavator (Fig. 1 Top Right) as well as its subparts (Bottom Right). This dense hierarchical 3D grouping enables applications such as 3D asset extraction and interactive segmentation.

GARField distills a set of 2D segmentation masks into a 3D volumetric scale-conditioned affinity field. Because grouping is an ambiguous task, these 2D labels can be overlapping or conflicting. These inconsistencies pose a challenge for distilling masks into consistent 3D groups. We ameliorate this issue by leveraging a *scale-conditioned* feature field. Specifically GARField optimizes a dense 3D feature field which is supervised such that feature distance reflects points' affinity. The scale conditioning enables two points to have higher affinity at a large scale but low affinity at a smaller scale (*i.e*. wedges of the same watermelon), as illustrated in Figure 2.

Though in principle GARField can distill any source of 2D masks, we derive mask candidates from Segment Anything Model (SAM) [15] because they align well with what humans consider as reasonable groups. We process input images with SAM to obtain a set of candidate segmentation masks. For each mask, we compute a physical scale based on the scene geometry. To train GARField, we distill candidate 2D masks with a contrastive loss based on mask membership, leveraging 3D scale to resolve inconsistencies between views or mask candidates.

A well-behaved affinity field has: 1) *transitivity* , which means if two points are mutually grouped with a third, they should themselves be grouped together, and 2) *containment*, which means if two points are grouped at a small scale, they should be grouped together at higher scales. GARField's use of contrastive loss in addition to a containment auxiliary loss encourages both of these properties.

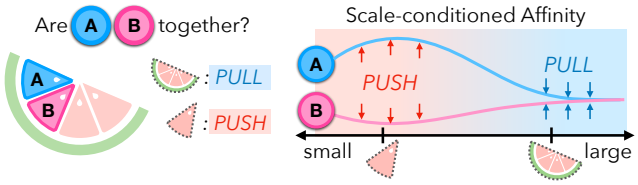

Figure 2. **Importance of Scale When Grouping** A single point may belong to multiple groups. GARField uses *scale-conditioning* to reconcile these conflicting signals into one affinity field.

With the optimized scale-conditioned affinity field, GARField extracts a 3D scene hierarchy via recursively clustering them at descending scales until no more clusters emerge. By construction, this recursive clustering ensures that generated groups are subparts of the prior cluster in a coarse-to-fine manner. We evaluate GARField on a variety of real scenes with annotated hierarchical groupings, testing its ability to capture object hierarchy, and its consistency across different views. By leveraging multiple views, GARField is able to produce detailed groupings, often improving upon the quality of input 2D segmentation masks. Moreover, these groups are 3D consistent by design, while 2D baselines do not guarantee view consistency. We show downstream applications of GARField for hierarchical 3D asset extraction and click-based interactive segmentation. Given GARField's scene decomposition capabilities, we're hopeful for its potential in other downstream applications like enabling robots to understand they can interact with or as a prior for dynamic reconstruction. See https://www.garfield.studio/ for code, data, and additional visualizations.

2. Related Work

Hierarchical Grouping Multi-level grouping has long been studied in 2D images since the early days of foreground segmentation [28]. Several methods build on this idea of spectral clustering for multi-level segmentation [5] and more complex hierarchical scene parsing [1, 25, 31]. These approaches rely on extracting contours either via classic texture cues and create a hierarchy either via a top-down [37] or bottom-up consolidation [1]. More recent deep learning approaches use edges [36] computed at multiple scales to create the hierarchy, and Ke *et al*. [11] proposes a transformer based unsupervised hierarchical segmentation approach guided by the outputs of a classic hierarchical segmentation [1].

Many works circumvent the question of ambiguity in grouping by defining a set of categories within which instances are to be segmented, *i.e*. panoptic segmentation [10, 14]. Recently, Segment Anything (SAM) [15] off-loads this ambiguity into prompting, where at each pixel multiple seg-

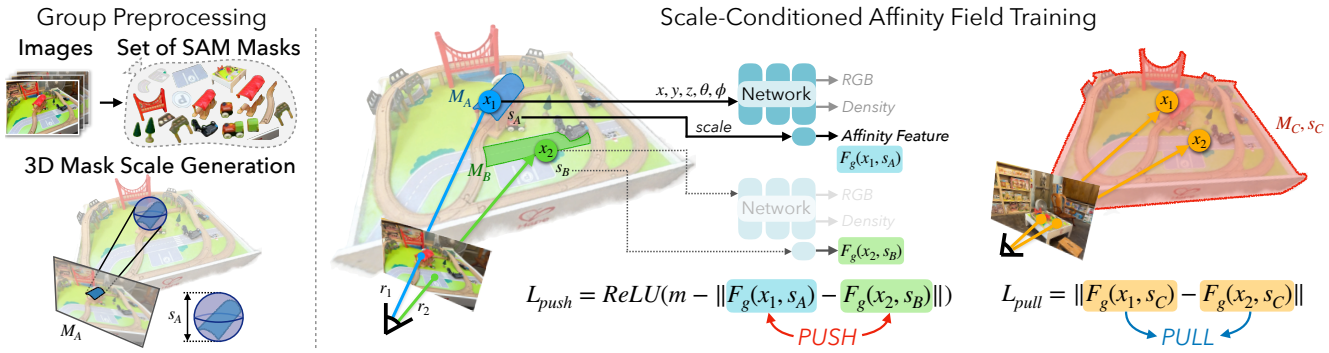

Figure 3. **GARField Method**: (Left) given an input image set, we extract a set of candidate groups by densely querying SAM, and assign each a physical scale by deprojecting depth from the NeRF. These scales are used to train a *scale-conditioned affinity field* (Right). During training, pairs of sampled rays are pushed apart if they reside in different masks, and pulled together if they land in the same mask. Affinity is supervised only at the scale of each mask, which helps resolve conflicts between them.

mentation masks can be proposed. However SAM does not recover a consistent set of hierarchical groups in the scene, which we enable by multi-scale 3D distillation.

Hierarchical part decomposition has also been explored in 3D objects, either in a supervised [17, 21, 35], or unsupervised manner [24]. Our approach distills information from a 2D model, and we consider full scenes while these approaches focus on 3D objects.

Segmentation in NeRFs Existing approaches for segmentation in NeRFs typically distill segmentation masks into 3D either by using ground-truth semantic labels [29, 38], matching instance masks [18], or training 3D segmentation networks on NeRF [34]. However, these techniques do not consider hierarchical grouping, and are only interested in a flat hierarchy of objects or instances. Ren et al. [27] leverages human interaction in the form of image scribbles to segment objects with interaction. More recently, Cen et al. [3] try to recover a 3D consistent mask from SAM by tracking the 2D masks between neighboring views via user prompting. Chen et al. [4] attempt this by distilling SAM encoder features into 3D and querying the decoder. In contrast with these approaches, our approach GARField does not require user input; it is able to obtain a hierarchical grouping of the scene automatically, and furthermore the recovered groups are view-consistent by definition.

3D Feature Fields Distilling higher-dimensional features into a neural field, in tandem with a radiance field (view-dependent color and density), has been thoroughly explored. Methods like Semantic NeRF [38], Distilled Feature Fields [16], Neural Feature Fusion Fields [33], and Panoptic Lifting [29] distill per-pixel 2D features into 3D by optimizing a 3D feature field to reconstruct the 2D features after volumetric rendering. These features can be either from pretrained vision models such as DINO or from semantic segmentation models. LERF [13] extends this idea to a scale-conditioned feature field, enabling the training of feature fields from global image embeddings like CLIP [26]. GARField similarly optimizes a scale-conditioned feature field in 3D; however, the purpose of the multi-scale features is to resolve ambiguity in grouping, instead of reconstructing an explicit 2D feature like CLIP. In addition LERF has no spatial grouping, a shortcoming GARField addresses. The aforementioned methods are based on direct supervision from image features, while other methods such as NeRF-SOS [8] and Contrastive Lift [2] optimize an arbitrary feature field at a single scale using a contrastive loss between pairs of rays based on similarity. GARField uses this contrastive approach because it allows for defining pairwise relationships between points based on mask labels. However, we design a scale-conditioned contrastive loss, which allows for distilling conflicting masks into 3D. In addition, GARField does not require the slow-fast formulation of Bhalgat et al. [2] for stable training, perhaps enabled by scale-conditioned training.

3. Method

3.1. 2D Mask Generation

GARField takes as input a set of posed images and produces a hierarchical 3D grouping of the scene, along with a standard 3D volumetric radiance field and a scale-conditioned affinity field. To do this, we first pre-process input images with SAM to obtain mask candidates. Next, we optimize a volumetric radiance field along with the affinity field which takes in a single 3D location and a euclidean scale, and outputs a feature vector. Affinity is obtained by comparing pairs of points' feature vectors. After optimization, the resulting affinity field can be used to decompose a scene by recursively clustering the feature embeddings in 3D at descending scales in a coarse-to-fine manner, or for segmenting user specified queries. The overall pipeline is illustrated in Figure 3.

In order to train a GARField, we first mine 2D mask candidates from an image and then assign a 3D scale for

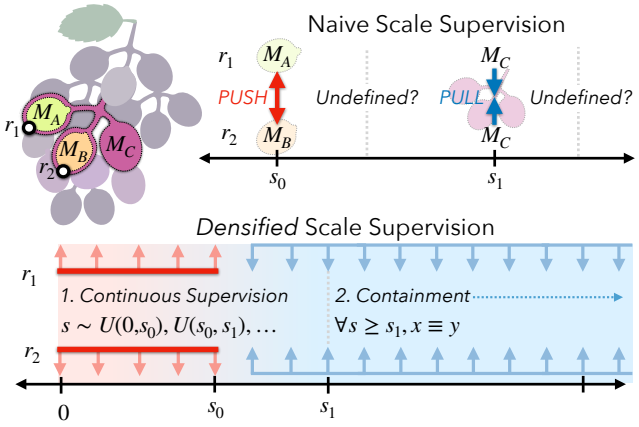

Figure 4. **Densified Scale Supervision**: Consider two grapes within a cluster. *Naively* using scale for contrastive loss supervises affinities only at the grape and grape trio levels, leaving entire intervals unsupervised. In GARField, we densify the supervision by 1) augmenting scale between mask euclidean scales and 2) imposing an auxiliary loss on containment of larger scales.

each mask. Specifically, we use SAM's automatic mask generator [15], which queries SAM in a grid of points and produces 3 candidate segmentation masks per query point. Then, it filters these masks by confidence and deduplicates nearly identical masks to produce a list of mask candidates of multiple sizes which can overlap or include each other. This process is done independently of viewpoint, producing masks which may not be consistent across views. In this work we aim to generate a hierarchy of groupings based on objects' physical size. As such, we assign each 2D mask a physical 3D scale as in Fig. 3. To do this we partially train a radiance field and render a depth image from each training camera pose. Next, for each mask we consider the 3D points within that mask x^i and set the 3D scale to $||2 * \texttt{std}(x^i)||$. This method ensures the 3D scale of masks resides in the same world-space, enabling scale-conditioned affinity.

3.2. Scale-Conditioned Affinity Field

Scale-conditioning is a key component of GARField which allows consolidating inconsistent 2D mask candidates: The same point may be grouped in several ways depending on the granularity of the groupings desired. Scale-conditioning alleviates this inconsistency because it resolves ambiguity over which group a query should belong to. Under scale-conditioning, conflicting masks of the same point no longer fight each other during training, but rather can coexist in the same scene at different affinity scales.

We define the scale-conditioned affinity field $F_g(x, s) \mapsto R^d$ over a 3D point x and euclidean scale s, similar to the formulation in LERF [13]. Output features are constrained to a unit hyper-sphere, and the affinity between two points at a scale is defined by $A(x_1, x_2, s) = -||F_g(x_1, s) - F_g(x_2, s)||_2$. These features can be volumetrically rendered

Figure 5. **3D Asset Extraction with Interactive Selection**: Users can interactively select view-consistent 3D groups with GARField using a click point and a scale.

with a weighted average using the same rendering weights based on NeRF density to obtain a value on a per-ray basis.

3.2.1 Contrastive Supervision

The field is supervised with a margin-based contrastive objective, following the definition provided by DrLIM [9]. There are two core components of the loss: at a given scale, one which pulls features within the same group to be close, and another which pushes features in different groups apart.

Specifically, consider two rays r_A, r_B sampled from masks $\mathcal{M}_A, \mathcal{M}_B$ within the same training image, with corresponding scales s_A and s_B. We can volumetrically render the scale-conditioned affinity features along each ray to obtain ray-level features F_A and F_B. If $\mathcal{M}_A = \mathcal{M}_B$, the features are pulled together with L2 distance: $\mathcal{L}_{\text{pull}} = ||F_A - F_B||$. If $\mathcal{M}_A \neq \mathcal{M}_B$, the features are pushed apart: $\mathcal{L}_{\text{push}} = \text{ReLU}(m - ||F_A - F_B||)$ where m is the lower bound distance, or margin. Importantly, this loss is only applied among rays sampled from the same image, since masks across different viewpoints have no correspondence.

3.2.2 Densifying Scale Supervision

The supervision provided by the previous contrastive losses alone are not sufficient to preserve hierarchy. For example in Fig. 10, although the egg is correctly grouped with the soup at scale 0.22, at a larger scale it fragments apart. We hypothesize this grouping instability is because 1) scale supervision is defined sparsely only when a mask exists and 2) nothing imposes containment such that small scale groups remain at larger scales. We address these shortcomings here by introducing the following modifications:

Continuous scale supervision By using 3D mask scales, groups are only defined at discrete values where masks are chosen. This results in large unsupervised regions of scale, as shown at the top of Fig. 4. We densify scale supervision by augmenting the scale s uniformly randomly between the current mask's scale and the next small-

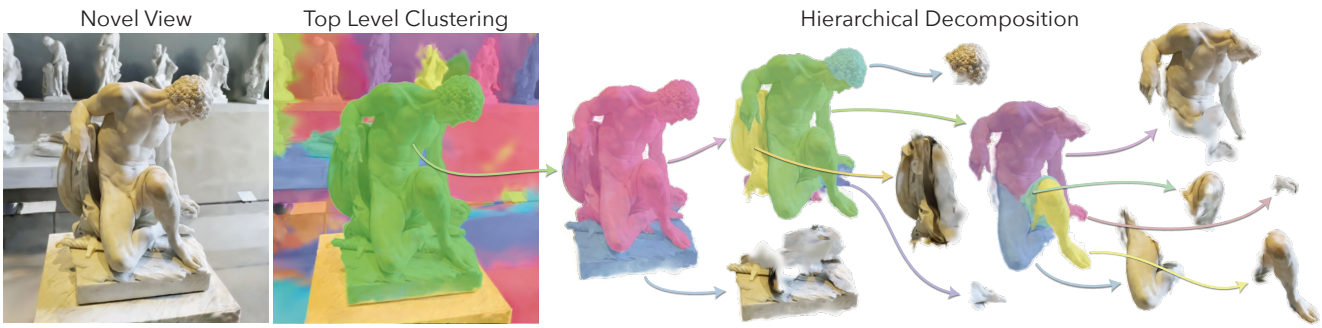

Figure 6. **3D Decomposition:** GARField can be recursively queried at decreasing scale to cluster a scene into objects and their subparts.

est mask's scale. When a ray's mask is the smallest mask for the given viewpoint, we interpolate between 0 and s_0. This ensures continuous scale supervision throughout the field leaving no unsupervised regions.

Containment Auxiliary Loss: If two rays r_1 and r_2 are in the same mask with scale s, then they should also be pulled together at any scale larger than s. Intuitively, two grapes within the same cluster (Fig. 4) are also grouped together at larger scales (e.g., the entire bunch). At each training step, for the rays grouped together at scale s, we additionally sample a larger scale $s' \sim \mathcal{U}(s, s_{max})$ at which the rays are also pulled together. This ensures that affinities at smaller scales are not lost at larger scales.

3.2.3 Ray and Mask Sampling

Just like standard NeRF training, we sample rays over which to compute losses. Because GARField uses a contrastive loss within each train image, naively sampling pixels uniformly during training is inadequate to provide a training signal in each minibatch of rays. To ensure sufficient pairs in each train batch, we first sample N images, and sample M rays within each image. To balance the number of images as well as the number of point pairs for supervision, we sample 16 images and 256 points per image, resulting in 4096 samples per train iteration.

For each ray sampled, we must also choose a mask to use as the group label for the train step in question. To do this, we retain a mapping from pixels to mask labels throughout training, and at each train step randomly select a mask for each ray from its corresponding list of masks. There are two important caveats in this sampling process: **1)** The probability a mask is chosen is weighted inversely with the log of the mask's 2D pixel area. This prevents large scales from dominating the sampling process, since larger masks can be chosen via more pixels. **2)** During mask selection we coordinate the random scale chosen across rays in the same image to increase the probability of positive pairs. To do this, we sample a single value between 0 and 1 per image, and index into each pixel's mask probability CDF with the same value, ensuring pixels which land within the same group are assigned the same mask. Otherwise, the loss is dominated by pushing forces which destabilize training.

3.3. Implementation Details

The method is built in Nerfstudio [32] on top of the Nerfacto model by defining a separate output head for the grouping field. The grouping field is represented with a hashgrid [23] with 24 layers and a feature dimension of 2 per layer, and a 4-layer MLP with 256 neurons and ReLU activation which takes in scale as an extra input concatenated with hashgrid feature. We cap scale at $2\times$ the extent of cameras, and normalize the scale input to the MLP using sklearn's quantile transform on the distribution of computed 3D mask scales (Sec 3.1). Output embeddings are $d = 256$ dimensions. Gradients from the affinity features do not affect the RGB outputs from NeRF, as these representations share no weights or gradients.

We begin training the grouping field after 2000 steps of NeRF optimization, giving geometry time to converge. In addition, to speed training we first volumetrically render the hash value, then use it as input to the MLP to obtain a ray feature. With this deferred rendering, the same ray can be queried at different scales with only one extra MLP call. Both volume-rendered (2D) and point-wise (3D) hashgrid values are normalized before inputting them to the MLP. Preprocessing SAM masks takes around 3-10 minutes, followed by about 20 minutes for training on a GTX 4090.

4. Hierarchical Decomposition

Once we have optimized a scale-conditioned affinity, GARField generates a hierarchy of 3D groups, organized in a tree such that each node is broken into potential subgroups. To do this we recursively cluster groups by decreasing the scale for affinity, using HDBSCAN [19], a density based clustering algorithm which does not require a prior on number of clusters. This clustering process can be done in 2D on volumetrically rendered features in an image which yields masks, or in 3D across points to yield pointclouds. See Fig. 6 for a visualization of scene decomposition.

Initialization: First, to initialize the top-level nodes in the

Figure 7. **Results**: From a GARField we extract objects from the global scene by selecting top-level clusters, then visualize their local clusters at decreasing scales. GARField can produce complete 3D object masks, and break these objects into meaningful subparts based on the input masks. We use Gaussian Splats [12] to produce these visualizations in 3D. See the website for video results.

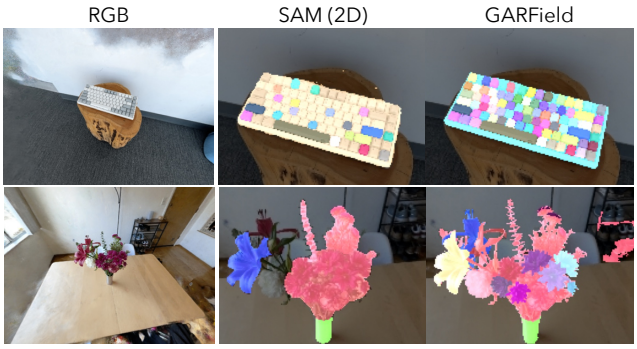

Figure 8. **Segment-Anything [15] vs. GARField**: SAM's automatic mask generator may struggle with recalling all masks, especially when there are clusters of small masks and the camera is far away from the object. In contrast, GARField's scale-conditioned affinity field incorporates masks from multiple viewpoints in 3D.

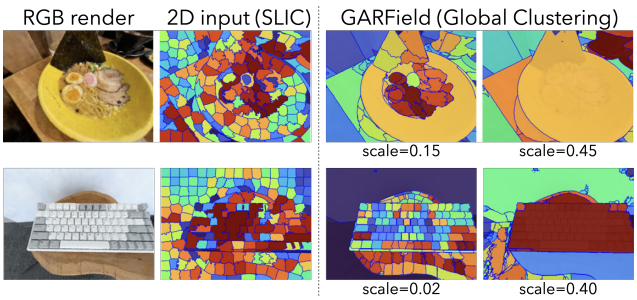

Figure 9. **GARField using SLIC as input**: Superpixel groups from varying viewpoints are reconciled into 3D hierarchy.

hierarchy, we first globally cluster features at a large scale $s_{\max}$, which we set to 1.0 for all experiments, corresponding to the extent of the input cameras' positions.

Recursive Clustering: Next, to produce a hierarchical tree of scene nodes, we iteratively reduce scale by a fixed epsilon (we use 0.05), running HDBSCAN on each leaf node. We use a relative resolution, instead of calculating a metric one (*e.g.* 5cm), to use across both small- and large-scale scenes. If HDBSCAN returns more than one cluster for a given node, we add those clusters as children and recurse. When all nodes reach scale 0, we return the current tree.

5. Experiments

We assess GARField's ability to decompose in-the-wild 3D scenes into hierarchical groups which vary widely in size and semantics. Existing 3D scan datasets tend to focus on object-level scans [7, 22], are simulated [2], or contain primarily indoor household scenes [6]. To evaluate GARField, we instead use a wide variety of indoor and outdoor scenes from the Nerfstudio and LERF datasets, as well as additional captures for this paper. We experiment on scenes which possess significant object hierarchy, testing the decomposition ability of GARField.

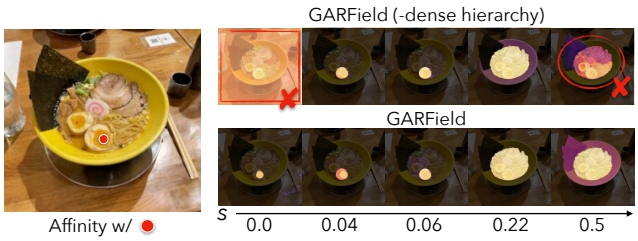

Figure 10. **Ablation**: Without dense hierarchy supervision, affinities may be inconsistent across scales, *e.g.* spurious large affinities at unsupervised scales, or unexpected drops at larger scales.

5.1. Qualitative Scene Decomposition

We use Gaussian Splatting [12] to visualize the decomposition by querying GARField's affinity field at gaussian centers. GS's explicit representation speeds up the process of clustering, segmenting, and manipulating a scene compared to NeRFs, which would require densely sampling points to cluster, then propagating those groups to the underlying field. See the Appendix for a full description of the pipeline. All renderings are of complete 3D models, not segmentations of 2D image views.

We visualize two types of hierarchical clustering results. In Fig. 7 scenes are globally clustered at hand-selected coarse scales, then from these scene-wide clusters we select groups corresponding to few objects and further decompose them into subgroups. We visualize clusters obtained at successively decreasing scales, increasing the granularity of groups. GARField achieves high-fidelity 3D groupings across a wide range of scenes and objects, from man-made objects – such as keyboards, where each key is considered a group at a small scale, to the parts of the NERF gun and the Lego bulldozer – to complex natural objects like plants, where it groups individual flowers as well as their petal and leaves. See Fig. 11 for scene-wide cluster visualizations.

In Fig. 6 we visualize a tree decomposition produced by the method described in Sec. 4. We first show the global clustering at a top level node, from which we select the central statue to illustrate the tree decomposition. Arrows denote children in the hierarchy, illustrating how the statues decomposes gradually all the way down to its hair, legs, torso, etc. See the Supplement for more tree visualizations.

5.2. Quantitative Hierarchy

We quantitatively evaluate our approach against annotated images using two metrics: the first measuring view consistency against annotations from multiple views and the second measuring recall of various hierarchical masks via mIOU against ground truth human annotations. See the Supplement for the full list of scenes and annotations.

3D Completeness: For downstream tasks it is useful for groups to correspond to complete 3D objects, for example groups that contain an entire object rather than just one of its

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.** We report mIOU of scene annotations for one point with up to three levels of hierarchy. SAM struggles to produce view-consistent fine groups compared to GARField.

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:** We report mIOU against human annotations of multi-scale groups of different objects.

sides. Though GARField always produces view-consistent groups by construction, it may not necessarily contain complete objects. We evaluate for completeness by checking that an entire 3D object is grouped together across a range of viewpoints. To do this, on 5 scenes we choose a 3D point to be projected into 3 different viewpoints, and label 3 corresponding view-consistent ground truth masks containing that point at coarse, medium, and fine levels. At these points we mine multiple masks from GARField across multiple scales at 0.05 increments, where at each scale a mask is obtained based on feature similarity thresholded at 0.9. We also compare against SAM by clicking the point in the image and taking all 3 masks. We report the maximum mIOU computed over all candidate masks for both methods.

Results are shown in Table 1. GARField produces more complete 3D masks than SAM across viewpoints, resulting in higher mIOU with multi-view human annotations of objects. This effect is especially apparent at the most granular level, like the keyboard keys from afar in Fig. 8. See supplement for more comparisons and groundtruth masks.

Hierarchical Grouping Recall: Here we measure GARField's ability to recall groups at multiple granularities. Across 5 scenes, we choose one *novel* viewpoint and label up to 3 groundtruth hierarchical groups for 1-2 objects. GARField outputs a hierarchy of masks as described in Section 4 by clustering image-space features. We compare against SAM's automatic mask generation by keeping all output masks. We ablate GARField in two ways: (-scale) removes scale-conditioning; and (-hierarchy) removes the densified supervision in Sec. 3.2.2.

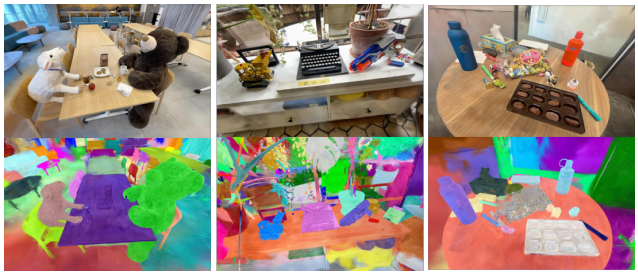

Figure 11. **Scene-Wide Clustering**: selected scenes from Fig. 7.

In Table 2 we report mIOU of the ground truth mask with the highest overlap, either from the set of SAM masks or the tree generated by GARField. Because GARField has fused groups from multiple perspectives, it results in higher fidelity groupings than any single view of SAM, leading to higher mIOU with annotations. Our ablations show that scale conditioning and scale densification is necessary for high quality groupings. Fig. 10 illustrates affinity degrading at higher scale with naive supervision.

6. Limitations

GARField at its core is distilling outputs from a 2D mask generator, so if the masks fail to contain a desired group, this will not emerge in 3D. We handle group ambiguity using physical size, but there could be multiple groupings within a single scale. For example, conflicts may happen with objects contained in a container because the container with and without the object can have the same scale. Future work could consider other ways to resolve grouping ambiguity such as affordances. Large groups in the background may fail due to scale clamping (s_{max} = 2× camera extent).

7. Conclusion

We present GARField, a method for distilling multi-level masks into a dense scale-conditioned affinity field for hierarchical 3D scene decomposition. By leveraging scale-conditioning, the affinity field can learn meaningful groups from conflicting 2D group inputs and break apart the scene at multiple different levels, which can be used for extracting assets at a multitude of granularities. GARField could have applications for tasks that require multi-level groupings like robotics, dynamic scene reconstruction, or scene editing.

8. Acknowledgements

This project was funded in part by NSF:CNS-2235013 and DARPA Contract No. HR001123C0021. CK and JK are supported in part by the NSF Graduate Research Fellowship Program under Grant No. DGE 2146752. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011. 2

[2] Yash Bhalgat, Iro Laina, João F Henriques, Andrew Zisserman, and Andrea Vedaldi. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. *NeurIPS*, 2023. 3, 7

[3] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Chen Yang, Wei Shen, Lingxi Xie, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with nerfs. 2023. 3

[4] Xiaokang Chen, jiaxiang Tang, Diwen Wan, Jingbo Wang, and Gang Zeng. Interactive segment anything nerf with feature imitation. *arXiv preprint arXiv:2211.12368*, 2023. 3

[5] T. Cour, F. Benezit, and J. Shi. Spectral segmentation with multiscale graph decomposition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 1124–1131 vol. 2, 2005. 2

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 7

[7] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 7

[8] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes. *arXiv preprint arXiv:2209.08776*, 2022. 3

[9] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, pages 1735–1742. IEEE, 2006. 4

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[11] Tsung-Wei Ke, Jyh-Jing Hwang, Yunhui Guo, Xudong Wang, and Stella X. Yu. Unsupervised hierarchical semantic segmentation with multiview cosegmentation and clustering transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 6, 7, 1

[13] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023. 3, 4

[14] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 2

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *ICCV*, 2023. 2, 4, 7, 8

[16] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *NeurIPS*, 35:23311–23330, 2022. 3

[17] Jun Li, Kai Xu, Siddhartha Chaudhuri, Ersin Yumer, Hao Zhang, and Leonidas Guibas. Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 3

[18] Yichen Liu, Benran Hu, Junkai Huang, Yu-Wing Tai, and Chi-Keung Tang. Instance neural radiacne field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[19] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017. 5

[20] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[21] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas J Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*, 2019. 3

[22] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019. 7

[23] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 5

[24] Despoina Paschalidou, Luc Van Gool, and Andreas Geiger. Learning unsupervised hierarchical part decomposition of 3d objects from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1060–1070, 2020. 3

[25] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):128–140, 2016. 2

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[27] Zhongzheng Ren, Aseem Agarwala, Bryan Russell, Alexander G Schwing, and Oliver Wang. Neural volumetric object selection. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 6133–6142, 2022. 3

[28] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. 2

[29] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. *arXiv preprint arXiv:2212.09802*, 2022. 3

[30] Piotr Skalski. Make Sense. https://github.com/SkalskiP/make-sense/, 2019. 3, 4

[31] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136, 2011. 2

[32] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, et al. Nerfstudio: A modular framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*, 2023. 5

[33] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural Feature Fusion Fields: 3D distillation of self-supervised 2D image representations. In *3DV*, 2022. 3

[34] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021. 3

[35] Yanzhen Wang, Kai Xu, Jun Li, Hao Zhang, Ariel Shamir, Ligang Liu, Zhiquan Cheng, and Yueshan Xiong. Symmetry hierarchy of man-made objects. In *Computer graphics forum*, pages 287–296. Wiley Online Library, 2011. 3

[36] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 2

[37] Stella X Yu. Segmentation using multiscale cues. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, pages I–I. IEEE, 2004. 2

[38] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 3