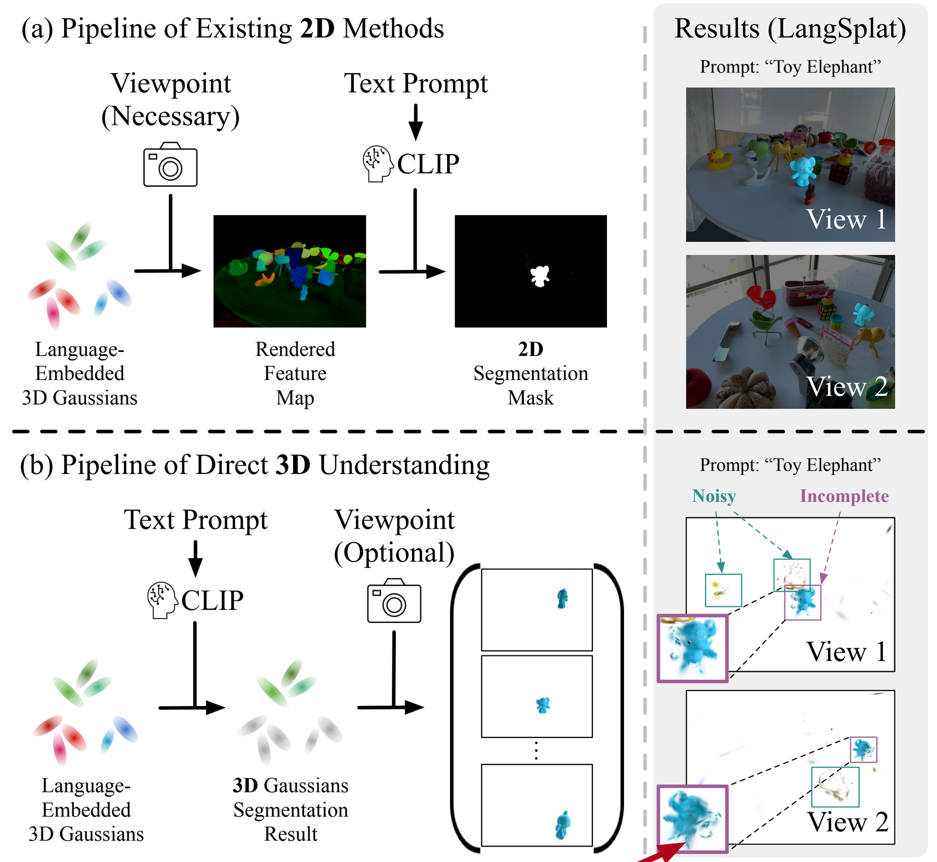
Tackling View-Dependent Semantics in 3D Language Gaussian Splatting

Jiazhong Cen[1], Xudong Zhou[1], Jiemin Fang[2], Changsong Wen[1], Lingxi Xie[2], Xiaopeng Zhang[2], Wei Shen[1], Qi Tian[2]

[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]Huawei Technologies Co., Ltd.

Introduction

Integrating vision-language features into 3D-GS is a common practice for open-vocabulary perception.

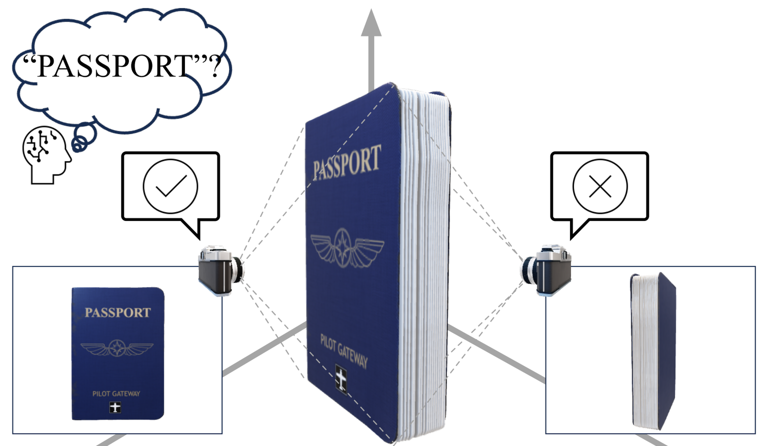


However, existing methods typically operate on rendered 2D feature maps and lack direct point-level understanding in 3D space.

But why?

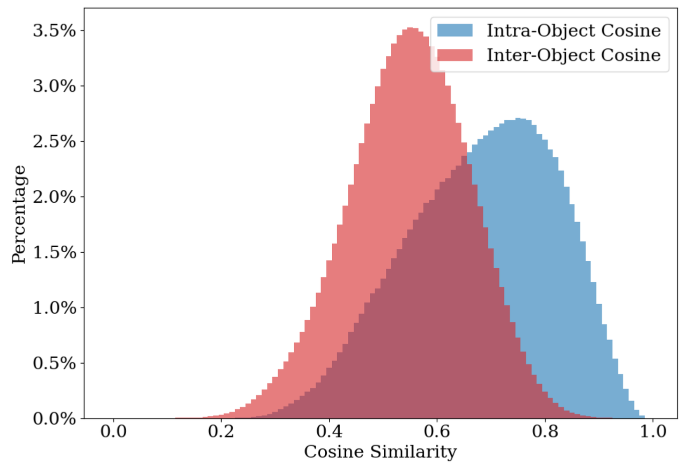
“横看成岭侧成峰
远近高低各不同”

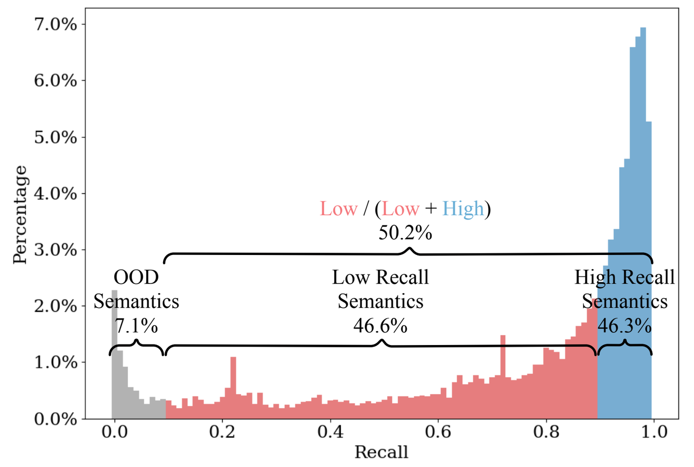
Consider a book, we can only recognize its name from limited viewpoints, indicating its semantics is view-dependent.



Existing methods ignore this characteristic, simply adopting 3D-GS to fit multi-view 2D semantic features! Thus leading to sub-optimal results.

Quantifying View-Dependency



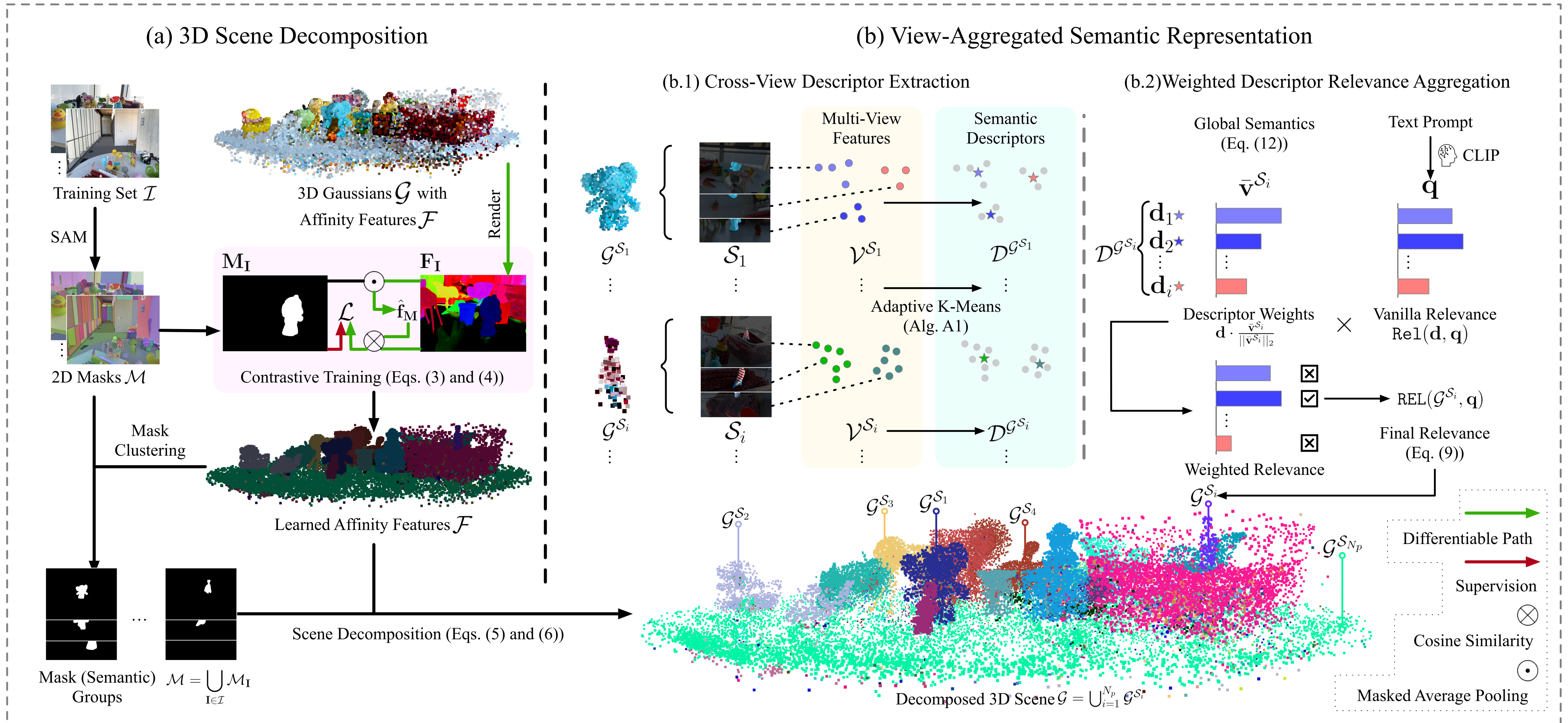


Cosine similarity of multi-view semantic features within the same object and across different objects: A substantial overlap between the inter- and intra-object distributions, suggesting inconsistent multi-view semantics of each object.

Using single view semantics to retrieve the corresponding object:

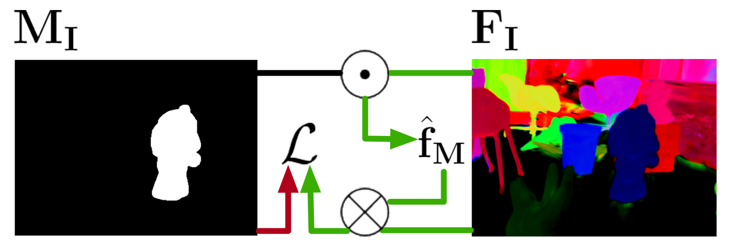
More than a half of semantic features can not completely retrieve its corresponding object.

Methodology



3D Scene Decomposition

LaGa first assign an affinity feature to each 3D Gaussian. During training, it renders the feature map $\mathbf{F_I}$



$$\hat{\mathbf{f}}_{\mathbf{M}^\mathbf{I}} = \mathrm{MAP}(\mathbf{M}^\mathbf{I}\mathbf{F_I}) = \frac{1}{\sum_{\mathbf{p}\in\delta(\mathbf{I})}\mathbf{M}^\mathbf{I}(\mathbf{p})} \sum_{\mathbf{p}\in\delta(I)} \mathbf{M}^\mathbf{I}(\mathbf{p})\mathbf{F}_I(\mathbf{p}).$$

$$\mathcal{L} = \sum_{\mathbf{I}\in\mathcal{I}} \sum_{\mathbf{M}\in\mathcal{M}^\mathbf{I}} \sum_{\mathbf{p}\in\delta(\mathbf{I})} (1-2\mathbf{M}(\mathbf{p}))\max\left(\langle \hat{\mathbf{f}}_\mathbf{M}, \mathbf{F_I}(\mathbf{p})\rangle, 0\right).$$

LaGa conducts clustering on the learned mask prototypes $\hat{\mathbf{f}}_\mathbf{M}$. The masks of each 3D object will be grouped into a same cluster $\mathcal{S}_i$

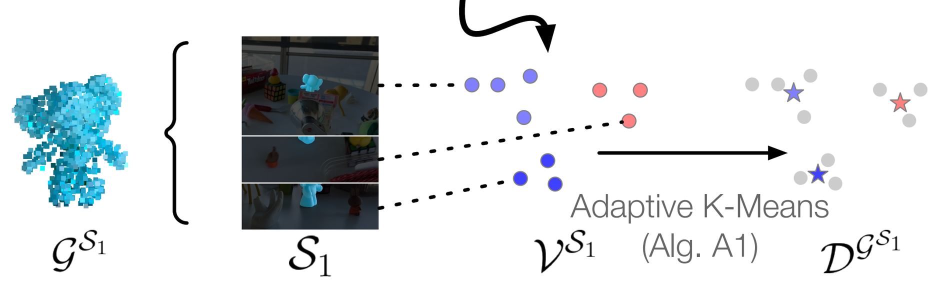
To decompose the 3D scene, LaGa derives classifier $\mathbf{t}^{\mathcal{S}_i}$ from the mask clusters. The 3D Gaussian with $\mathbf{f_g}$ belongs to the i^*-th object.

$$\mathbf{t}^{\mathcal{S}_i} = \frac{1}{|\mathcal{S}_i|}\sum_{\mathbf{M}\in\mathcal{S}_i} \hat{\mathbf{f}}_\mathbf{M}. \qquad i^* = \arg\max_i \langle \mathbf{f_g}, \mathbf{t}^{\mathcal{S}_i}\rangle.$$

Cross-View Descriptor Extraction

For each 3D object, LaGa uses CLIP to extract its multi-view semantics:

$$\mathcal{V}^{\mathcal{S}_i} = \left\{\mathbf{v}^\mathbf{M} \mid \mathbf{M}\in\mathcal{S}_i\right\}.$$



The multi-view semantics of different 3D objects exhibit varying levels of complexity.

To account for this, LaGa employs an **adaptive** K-Means algorithm (Alg. A1) on each object's semantic features to adaptively extract a set of descriptors:

$$\mathcal{D}^{\mathcal{G}^{S_i}} = \left\{\mathbf{d}_i \in \mathbb{R}^{C'} \,\middle|\, i\in\left\{1,\ldots,N^{\mathcal{G}^{S_i}}\right\}\right\}.$$

Note: C denotes the number of channels of affinity features and C' for the CLIP semantic features.

Weighted Descriptor Relevance Aggregation

After assigning a set of descriptors to each 3D object, LaGa adjust the importance of descriptors according to two metrics:

1. Directional Consistency: reliable descriptors should have higher cosine similarity with the objects global feature.
2. Internal Compactness: good descriptors should be generated by compact multi-view l2-normalized features. Its l2-norm indicates this compactness.

$$\omega^\mathbf{d} = \mathbf{d}\cdot\frac{\bar{\mathbf{v}}^{\mathcal{S}_i}}{||\bar{\mathbf{v}}^{\mathcal{S}_i}||_2} = \underbrace{\frac{\mathbf{d}}{||\mathbf{d}||_2}\cdot\frac{\bar{\mathbf{v}}^{\mathcal{S}_i}}{||\bar{\mathbf{v}}^{\mathcal{S}_i}||_2}}_{\text{(i) Directional Consistency}} \times \underbrace{||\mathbf{d}||_2}_{\text{(ii) Internal Compactness}}$$

During inference, given the text prompt $\mathbf{q}$, the highest weighted relevance score of the object is regarded as its final score:

$$\mathtt{REL}(\mathcal{G}^{\mathcal{S}_i},\mathbf{q}) = \max_{\mathbf{d}\in\mathcal{D}^{\mathcal{G}^{\mathcal{S}_i}}} \omega^\mathbf{d}\cdot\mathtt{Rel}(\mathbf{d},\mathbf{q}).$$

Quantitative Results

Results on the LERF-OVS dataset
(* denotes concurrent preprints)

	Methods	F.	T.	R.	W.	Mean
2D	LSeg	7.6	21.7	7.0	29.9	16.6
	LERF	38.6	45.0	28.2	37.9	37.4
	LEGaussians	60.3	44.5	52.6	41.4	46.9
	LangSplat	44.7	65.1	51.2	44.5	51.4
	N2F2	47.0	69.2	56.6	47.9	54.4
	OccamLGS*	58.6	70.2	51.0	65.3	61.3
	VLGS*	58.1	73.5	61.4	54.8	62.0
3D	OpenGaussian†	39.3	60.4	31.0	22.7	38.4
	SAGA‡	36.2	19.3	53.1	14.4	30.7
	LangSplat‡	25.9	35.6	29.3	33.5	31.1
	LEGaussians‡	31.2	34.5	17.6	17.3	25.2
	OpenGaussian‡	61.1	59.1	29.2	31.9	45.3
	SuperGSeg*	43.7	55.3	18.1	26.7	35.9
	LaGa (ours)	**64.1**	**70.9**	**55.6**	**65.6**	**64.0**

'2D' denotes conducting segmentation on rendered feature map. '3D' denotes conduct 3D segmentation then rendering the results to different views

Under the same experiment setting, LaGa surpasses previous method for **+18.7%** mIoU.

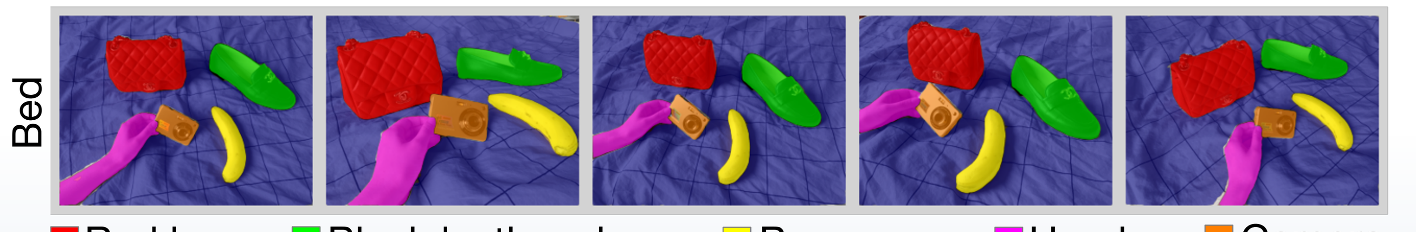
Qualitative Results

Results on the LERF-OVS dataset

Text Prompts: "Miffy" "Spatula" "Pirate hat" "Porcelain hand"

Text Prompts: "Three cookies" "Tea in a glass" "Paper napkin" "Apple"

Results on the 3D-OVS dataset



Red bag · Black leather shoe · Banana · Hand · Camera · White sheet

Please refer to our paper and project page for more visualization!

Conclusion

This paper investigate the **view-dependency of multi-view semantics** in 3D objects—an issue ignored by prior methods. We propose **LaGa**, which **decomposes scenes into 3D objects** and **gathers their multi-view semantics adaptively**. Extensive experiments demonstrate its effectiveness.