Open Vocabulary 3D Scene Understanding

Muhammad Aman Ahmad Tifli

Example task

“Retrieve Baymax from the top of the table”

Open Vocabulary 3D Scene Understanding

Muhammad Aman Ahmad Tifli

Definition

Open Vocabulary

- Beyond Predefined Categories
- Handling Previously Unseen Objects
- Free-form language Integration
- Contextual Understanding
- **Key-word:** Zero-shot

3D Scene Understanding

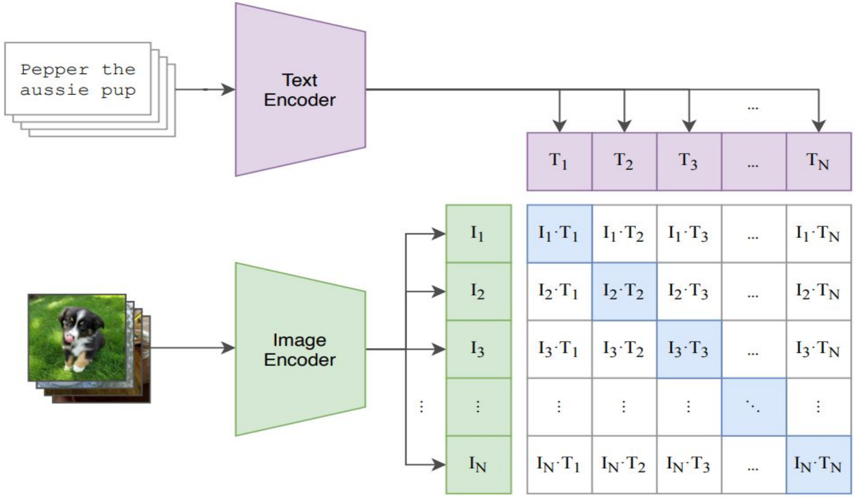
- Given images and 3D point clouds of an environment
- Goal: Semantic understanding of the environment for flexible robot applications

CLIP (*Constrastive Language-Image Pre-Training*)

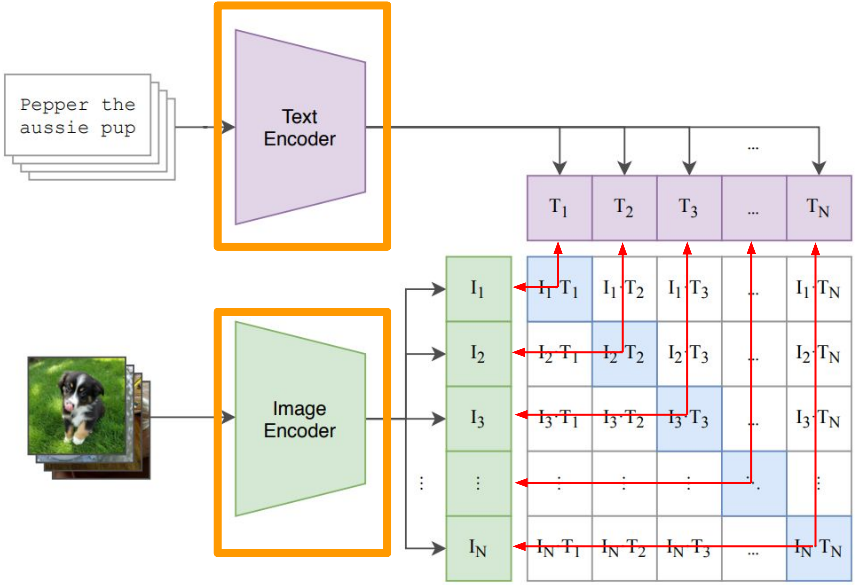
OpenAI model trained to predict similarities between text prompts and images

Trained on 400 million image-text pairs

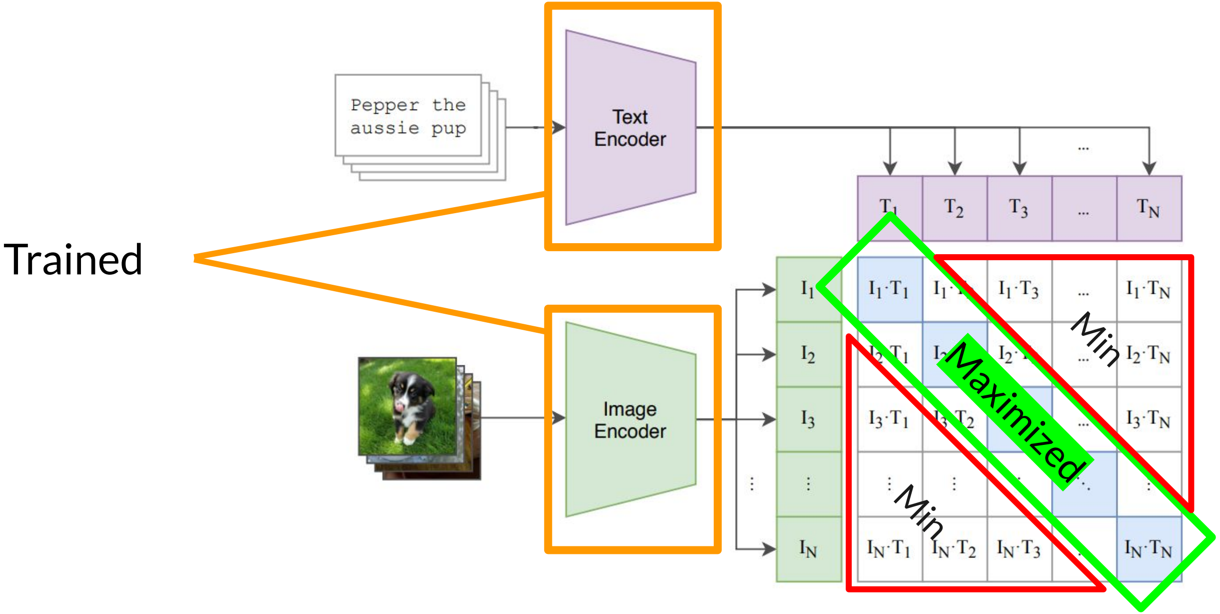
Diverse visual concepts learned from natural language



CLIP Embeddings Generation



CLIP Training



*Contrastive language-image pre-training!

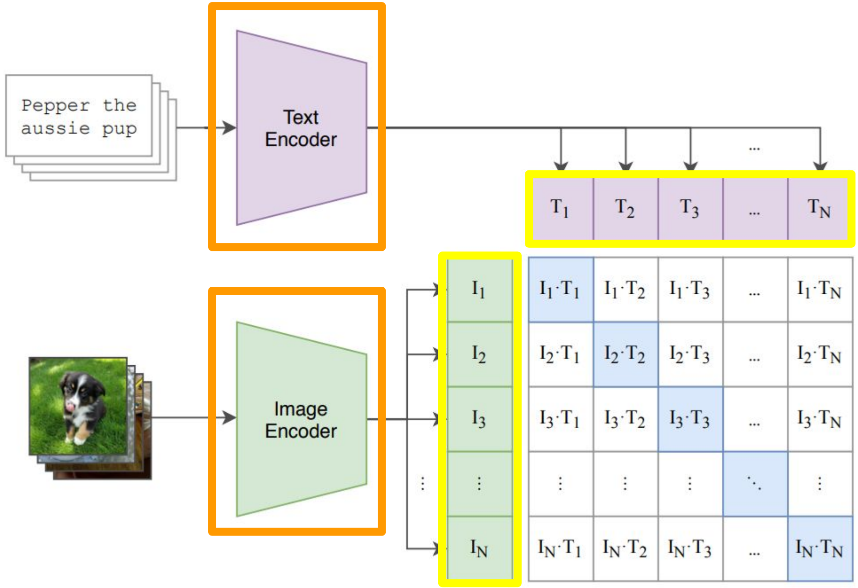
Introducing CLIP

OpenAI model trained to predict similarities between text prompts and images

Trained on 400 million image-text pairs

Diverse visual concepts learned from natural language

Encoders can be reused to create image and text embeddings



Relevant Papers Overview

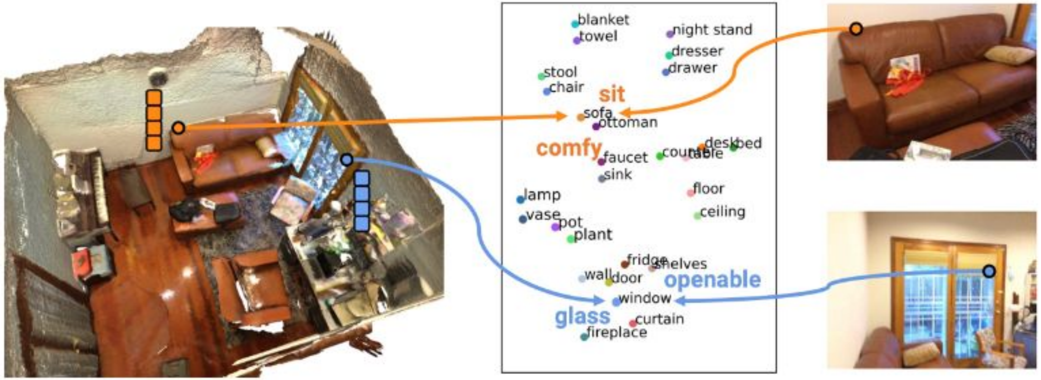
ConceptFusion: Open-Set Multimodal 3D Mapping

OpenScene: 3D Scene Understanding with Open Vocabularies

OpenShape: Scaling Up 3D Shape Representation Towards Open-World Understanding

Relevant Papers Overview

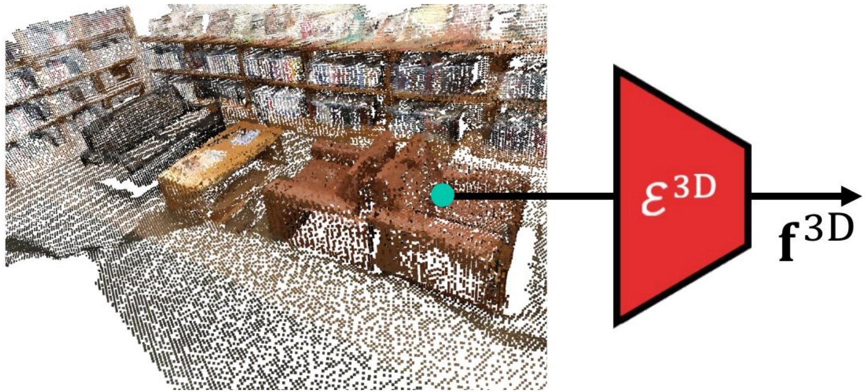
- ConceptFusion
- OpenScene
- OpenShape



Each 3D point matched with CLIP features from corresponding pixels in 2D images

Relevant Papers Overview

- ConceptFusion
- OpenScene
- OpenShape



Training an encoder to generate CLIP embeddings for 3D points

ConceptFusion: Open-Set Multimodal 3D Mapping

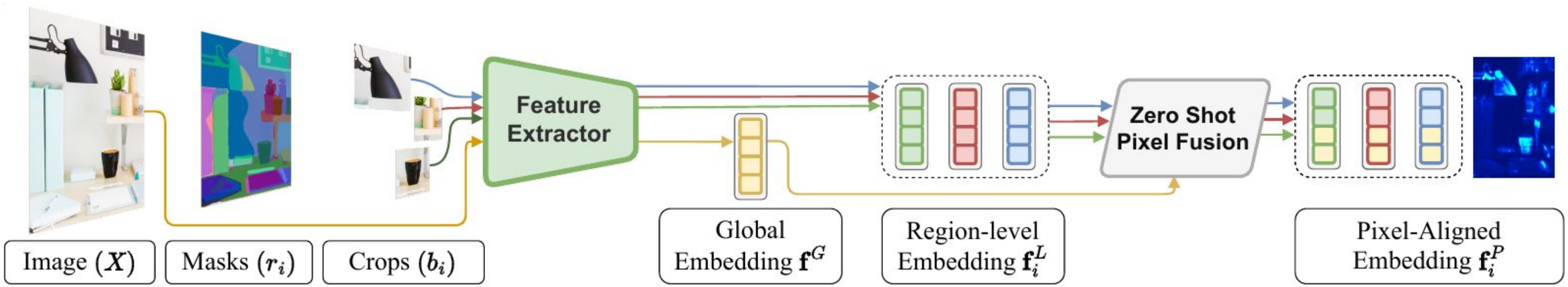
Jatavallbhula, Krishna Murthy, et al.

ConceptFusion: Open-Set Multimodal 3D Mapping
Method: Map representation

- Goal: A 3D-Map, $\mathbf{M}$, where each point has:
  - A vertex position, $\mathbf{v}_k \in \mathbb{R}^3$
  - A normal vector, $\mathbf{n}_k \in \mathbb{R}^3$
  - A confidence count, $\mathbf{c}_k$
  - A concept embedding $\mathbf{f}^P_k$

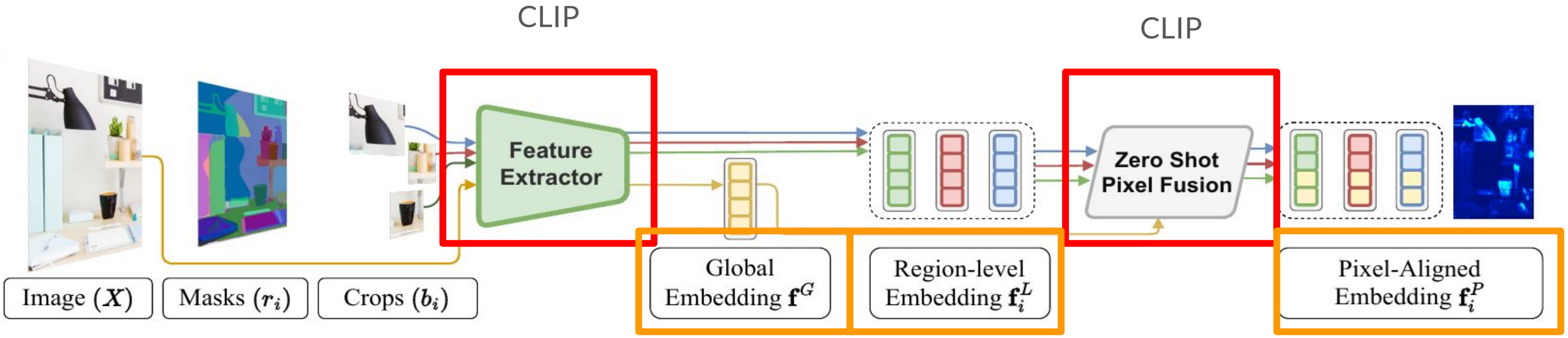
ConceptFusion: Open-Set Multimodal 3D Mapping

Method: Computing pixel-aligned features



ConceptFusion: Open-Set Multimodal 3D Mapping

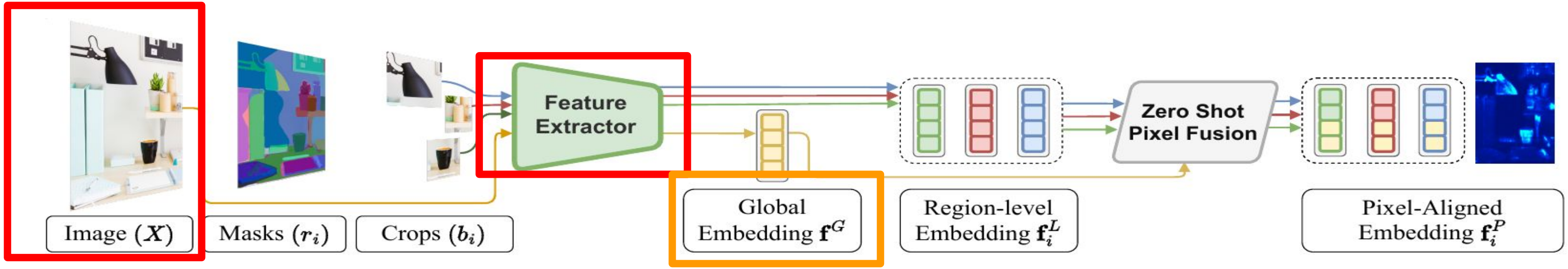
Image encoder as feature extractors



ConceptFusion: Open-Set Multimodal 3D Mapping

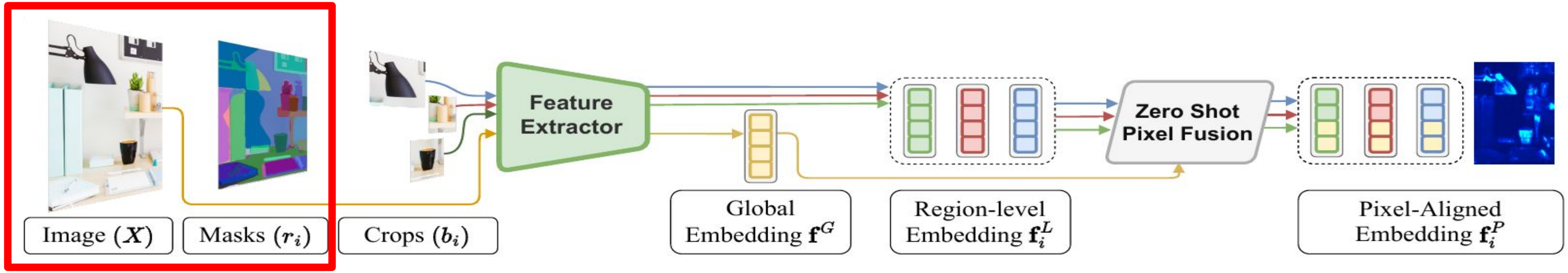
Global Embedding

$$f^G = F(X)$$



ConceptFusion: Open-Set Multimodal 3D Mapping

Region-level embedding

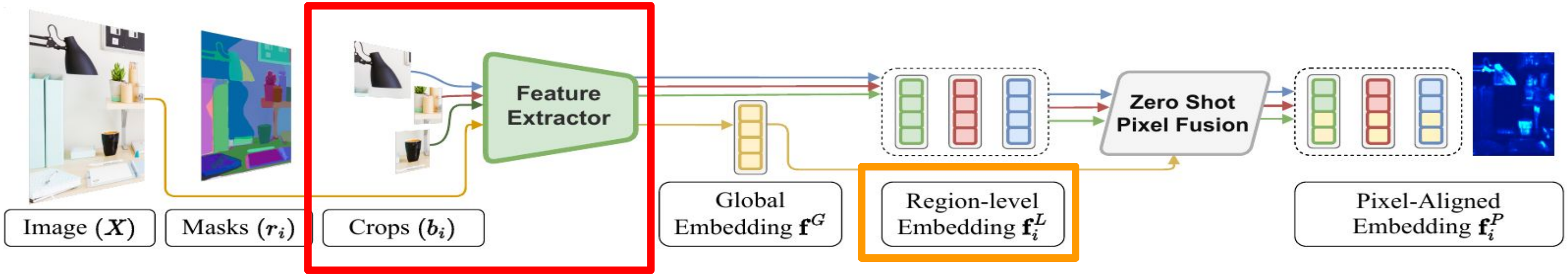


Segmented

(Mask2Former, SAM)

ConceptFusion: Open-Set Multimodal 3D Mapping

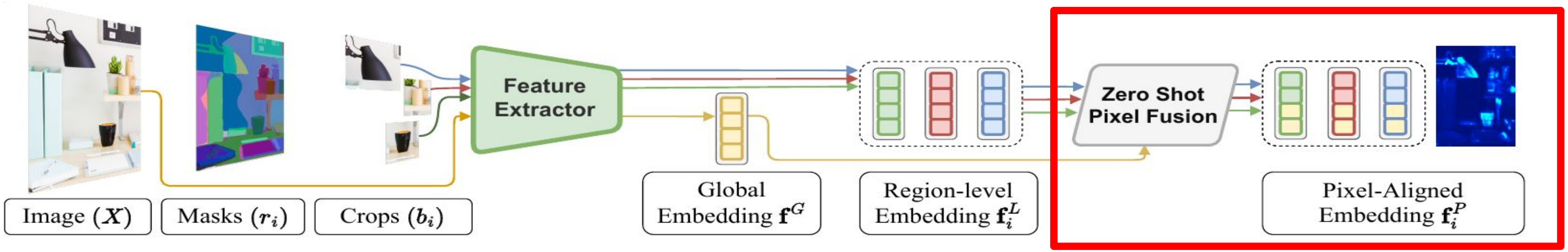
Region-level embedding



$$f_i^L = F(b_i)$$

ConceptFusion: Open-Set Multimodal 3D Mapping

Pixel Aligned embedding



$$\phi = \langle f_i^L, f^G \rangle$$

Cosine similarity

$$\varphi_{ij} = \langle f_i^L, f_j^L \rangle \qquad \bar{\varphi_{ij}} = \frac{1}{R}\Sigma_{j=1, j\neq 1}^{R} \varphi_{i,j}$$

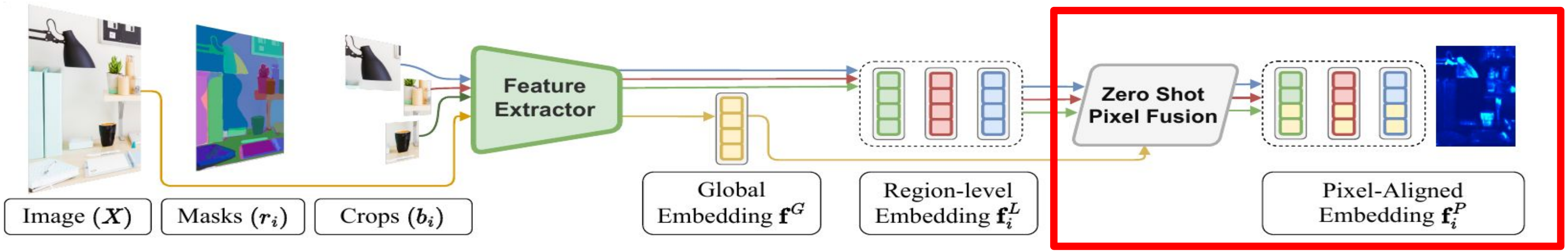
Uniqueness (similarity vs all others)

$$w_i = \frac{exp(\frac{\phi_i+\bar{\varphi_i}}{\tau})}{\Sigma_{i=1}^{R} exp(\frac{\phi_i+\bar{\varphi_i}}{\tau})}, \tau = 1$$

Mixing weight (Softmax)

ConceptFusion: Open-Set Multimodal 3D Mapping

Region-level embedding

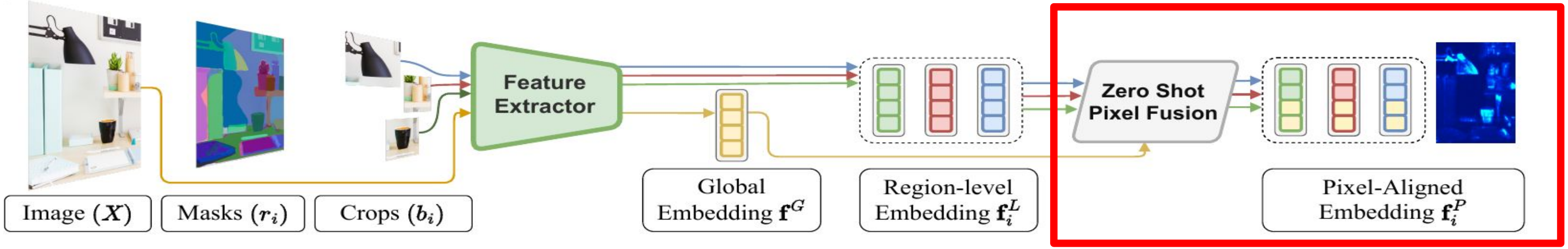


$$f_i^P = w_i f^G + (1 - w_i) f^L$$

Weighted combination
(Pixel-Aligned Embedding)

ConceptFusion: Open-Set Multimodal 3D Mapping

Region-level embedding



$$f_i^P \xrightarrow{\text{Mapped to pixel (u,v) in } r_i} f_{u,v}^P \xrightarrow{\text{Normalized after calculated for all regions in images in } X_t} f_{u,v,t}^P$$

Pixel-Aligned Embedding

Semantic context embedding

ConceptFusion: Open-Set Multimodal 3D Mapping

Fusing embeddings into the map

- Vertex and normal maps are first mapped to the global map using the camera pose
- For each pixel $(\mathbf{u},\mathbf{v})_t$ in image $\mathbf{X}_t$ that has a corresponding point $\mathbf{p}_k$ in global map, **M** the following is used:

$$f_{k,t}^{P} \leftarrow \frac{\bar{c_k} f_{k,t-1}^{P} + \alpha f_{u,v,t}^{P}}{\bar{c_k} + \alpha}$$

- Confidence, c_k, depends on distance to camera

ConceptFusion: Open-Set Multimodal 3D Mapping
Final map

- A 3D-Map, $\mathbf{M}$, where each point has:
    - A vertex position, $\mathbf{v}_k \in \mathbb{R}^3$
    - A normal vector, $\mathbf{n}_k \in \mathbb{R}^3$
    - A confidence count, $\mathbf{c}_k$
    - A concept embedding $\mathbf{f}^{P}_{k}$

ConceptFusion: Open-Set Multimodal 3D Mapping

Text Inference

"A comfy place to sit and watch TV" → Text Encoder → f_{text}

$$s_k = \langle f_{text}, f_k^P \rangle$$

Cosine similarity inference

Experiments & Results

ConceptFusion: Open-Set Multimodal 3D Mapping

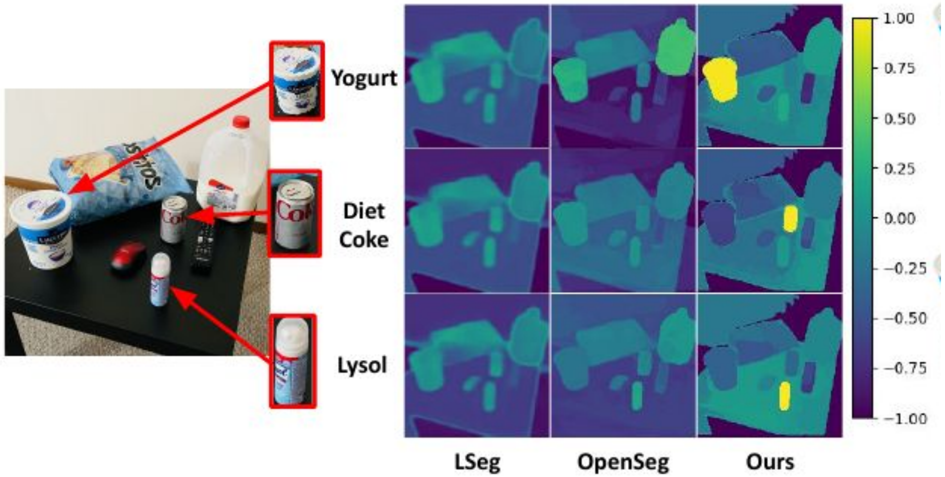
ConceptFusion: Open-Set Multimodal 3D Mapping
Experiments and results: Queries on UnCoCo dataset

- Evaluation of text-query-based object localization on UnCoCo dataset
  - **Given**: Text query i.e "lamp"
  - **Measured:** How much the predicted area matches the ground truth (IoU)
  - Evaluated against LSeg-3D, OpenSeg-3D (Supervised) and MaskCLIP-3D (Zero-shot)

		3D mIoU	IoU >0.15	IoU >0.25	IoU >0.5
Supervised	LSeg-3D	0.128	25%	16.66%	9.72%
	OpenSeg-3D	0.289	43.05%	36.11%	27.78%
	MaskCLIP-3D	0.091	25.97%	9.09%	1.30%
	ConceptFusion	**0.446**	**77.78%**	**69.44%**	**45.83%**

ConceptFusion: Open-Set Multimodal 3D Mapping
Concepts lost through fine-tuning

- Pixel-aligned embeddings capture fine-grained concepts
- Other approaches like LSeg and OpenSeg tend to forget such concepts through fine-tuning



ConceptFusion: Open-Set Multimodal 3D Mapping
Experiments and results: Real robotic systems

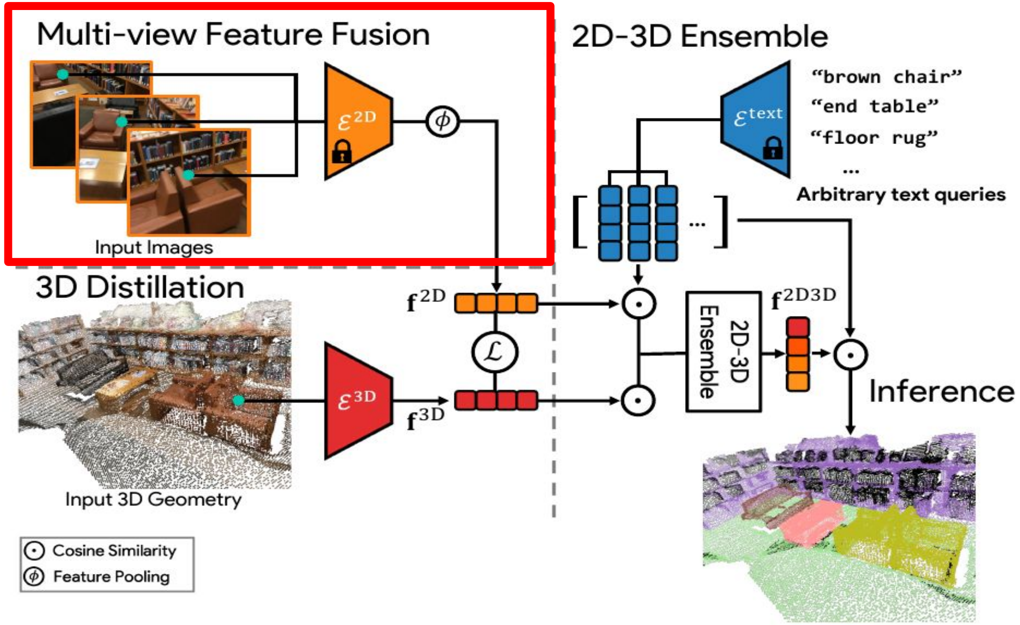
- Zero-shot tabletop rearrangement task
- Previously unseen objects
- Workspace sides tagged as left and right
- Goal: “Move Baymax to the right”

OpenScene: 3D Scene Understanding with Open Vocabularies

Peng, Songyou, et. al.

OpenScene: 3D Scene Understanding With Open Vocabularies
Method Overview

Similar to ConceptFusion



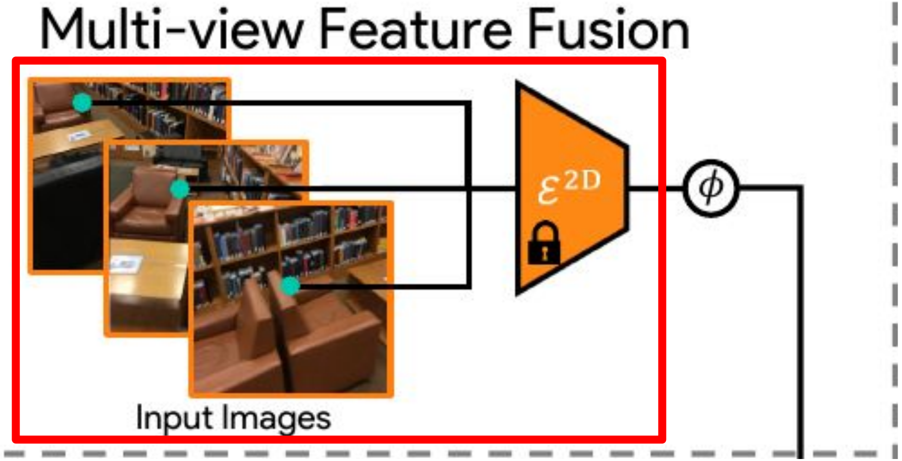
OpenScene: 3D Scene Understanding With Open Vocabularies
Method: Compute Per-Pixel features

Per-pixel features (LSeg, MSeg) $f_n = \epsilon^{2D}(X_n)$

For each 3D point, $\mathbf{p}$, in the point cloud, $\mathbf{P}$, the corresponding pixel, $\mathbf{u}$, of each input frame is calculated:

$$\tilde{u} = I_i \cdot E_i \cdot \tilde{p}$$

I_i, E_i: intrinsic and extrinsic matrices of the i-th frame



OpenScene: 3D Scene Understanding With Open Vocabularies
Method: Compute Per-Pixel features

Assume **K** associated pixels per 3D point, **p.**

All 2D embeddings fused by average pooling:

$$f^{2D} = \phi(f_1, \ldots, f_k)$$

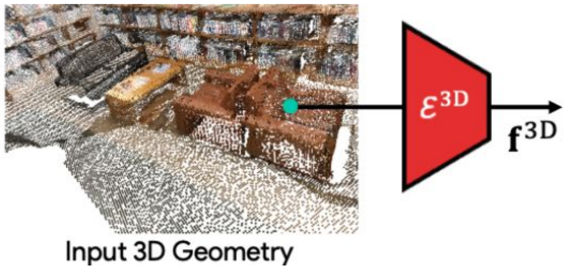
Repeated for all points in point cloud

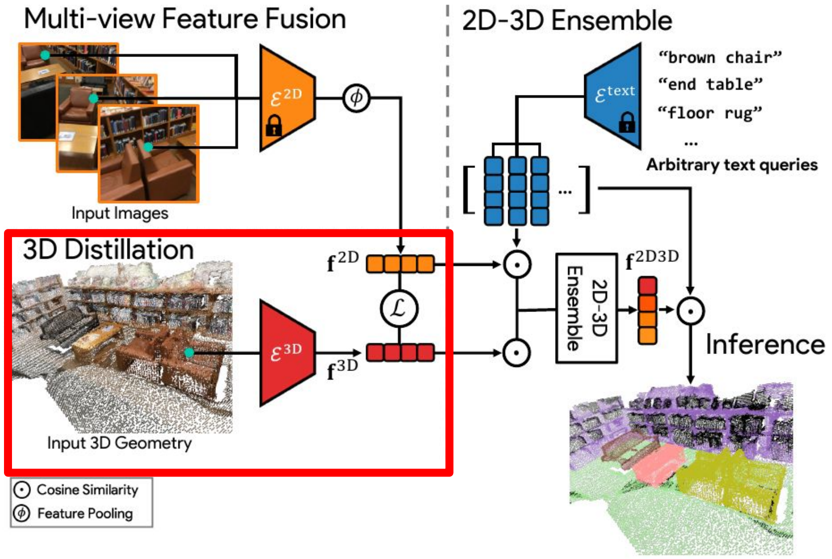
$$F^{2D} = \{f_1^{2D}, \ldots, f_k^{2D}\}$$

OpenScene: 3D Scene Understanding With Open Vocabularies
Distillation of 3D point network

$\mathbf{F}^{2D}$ can be inconsistent depending on the input image frames.

Solution: Distill a 3D point network (enocder)





3D Scene Understanding With Open Vocabularies

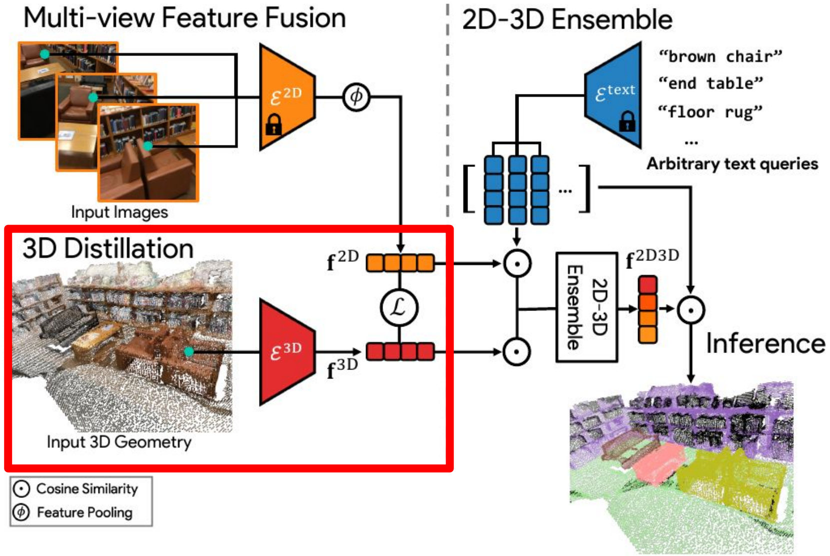
Distillation of 3D point network

MinkowskiNet as 3D-Semantic Segmentation backbone:

$$F^{3D} = \epsilon^{3D}(P)$$

Loss used to learn to create embeddings in **F^{2D}** space:

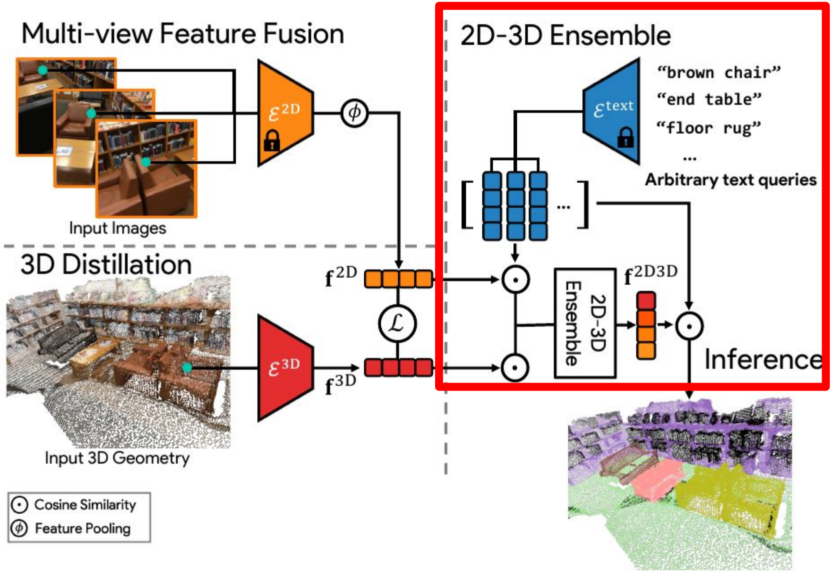
$$\mathcal{L} = 1 - cos(F^{2D}, F^{3D})$$



3D Scene Understanding With Open Vocabularies

2D-3D Feature Ensemble

- **Observation:**
  - **2D Features:** better for small objects and objects with ambiguous geometry
  - **3D Features:** better for objects with distinct shapes
- **Idea:** Combine both features



3D Scene Understanding With Open Vocabularies

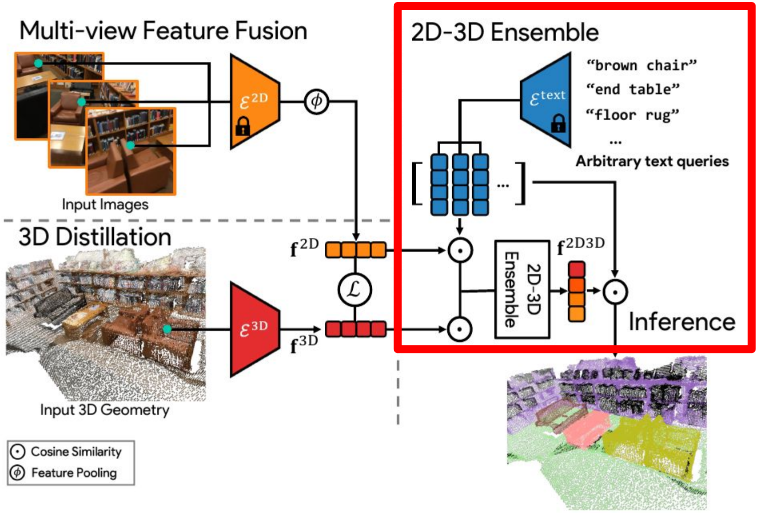
2D-3D Feature Ensemble

1. Text prompts (arbitrary of targeted) are provided and encoded with CLIP's text encoder
2. Cosine similarity of text embeddings calculated for all 2D and 3D features

$$s_n^{2D} = cos(f^{2D}, t_n) \qquad s_n^{3D} = cos(f^{3D}, t_n)$$

3. Maximum calculated and final feature f^{2D3D} has the highest score

$$s^{2D} = \max_n(s_n^{2D}) \qquad s^{3D} = \max_n(s_n^{3D})$$

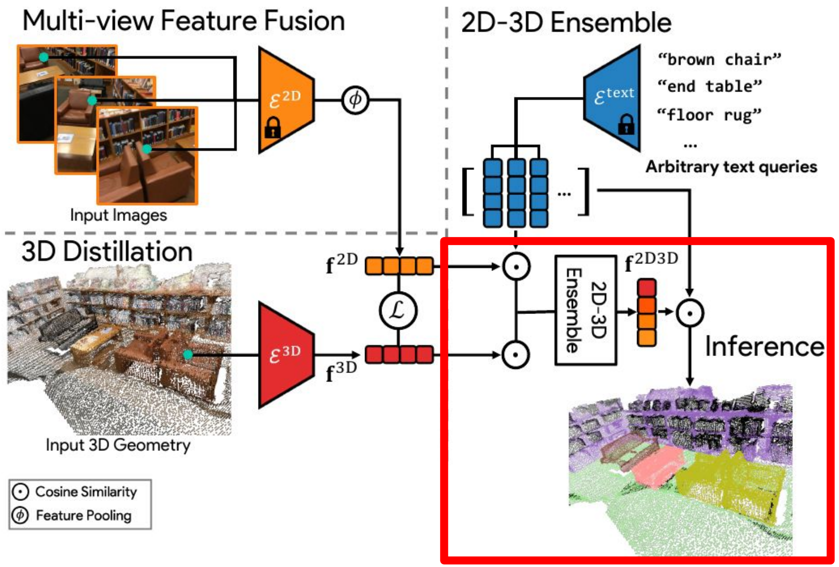


3D Scene Understanding With Open Vocabularies

Inference for Semantic Segmentation

Cosine similarity score between any of the previously discussed features, $\mathbf{f}^{2D}$, $\mathbf{f}^{3D}$, $\mathbf{f}^{2D3D}$, can be used for inference.

$$\arg\max_n \{cos(f^{2D3D}, t_n)\}$$



Experiments & Results

OpenScene: 3D Scene Understanding With Open Vocabularies

OpenScene: 3D Scene Understanding With Open Vocabularies

Comparison on zero-shot 3D semantic segmentation benchmarks

- Competitive even against fully-supervised approaches

	nuScenes [3]		ScanNet [11]		Matterport [4]	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
Fully-supervised methods						
TangentConv [51]	-	-	40.9	-	-	46.8
TextureNet [24]	-	-	54.8	-	-	63.0
ScanComplete [12]	-	-	56.6	-	-	44.9
DCM-Net [48]	-	-	65.8	-	-	66.2
Mix3D [40]	-	-	**73.6**	-	-	-
VMNet [22]	-	-	73.2	-	-	**67.2**
LidarMultiNet [60]	**82.0**	-	-	-	-	-
MinkowskiNet [10]	78.0	83.7	69.0	77.5	54.2	64.6
Zero-shot methods						
MSeg [29] Voting	31.0	36.9	45.6	54.4	33.4	39.0
Ours - LSeg	36.7	42.7	**54.2**	66.6	**43.4**	53.5
Ours - OpenSeg	**42.1**	**61.8**	47.5	**70.7**	42.6	**59.2**

OpenScene: 3D Scene Understanding With Open Vocabularies

Comparison on zero-shot 3D semantic segmentation benchmarks

- Competitive even against fully-supervised approaches
- Came closest to fully-supervised approaches on the Matterport dataset
- Matterport is the most diverse dataset (harder to train)

	nuScenes [3]		ScanNet [11]		Matterport [4]	
	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
Fully-supervised methods						
TangentConv [51]	-	-	40.9	-	-	46.8
TextureNet [24]	-	-	54.8	-	-	63.0
ScanComplete [12]	-	-	56.6	-	-	44.9
DCM-Net [48]	-	-	65.8	-	-	66.2
Mix3D [40]	-	-	**73.6**	-	-	-
VMNet [22]	-	-	73.2	-	-	**67.2**
LidarMultiNet [60]	**82.0**	-	-	-	-	-
MinkowskiNet [10]	78.0	83.7	69.0	77.5	54.2	64.6
Zero-shot methods						
MSeg [29] Voting	31.0	36.9	45.6	54.4	33.4	39.0
Ours - LSeg	36.7	42.7	**54.2**	66.6	**43.4**	53.5
Ours - OpenSeg	**42.1**	**61.8**	47.5	**70.7**	42.6	**59.2**

OpenScene: 3D Scene Understanding With Open Vocabularies

Impact of increasing number of object classes

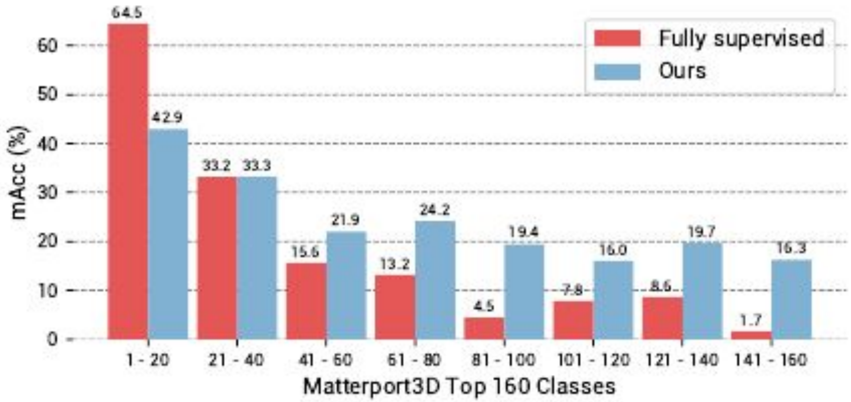
Dataset split into most frequent K classes, where K = 21, 40, 80, 160

A different MinkoswkiNet was trained for each K, while OpenScene was always kept the same

OpenScene outperforms the fully supervised method as classes increase and instances per class decrease

	$K = 21$	$K = 40$	$K = 80$	$K = 160$
Fully-supervision [10]	**64.5**	50.8	33.4	18.4
Ours	59.2	**50.9**	**34.6**	**23.1**

(a) Results on different number of classes in mAcc



OpenShape: Scaling Up 3D Shape Representation Towards Open-World Understanding

Liu, Minghua, et al.

OpenShape: Scaling up 3D Shape Representation Towards Open-World Understanding

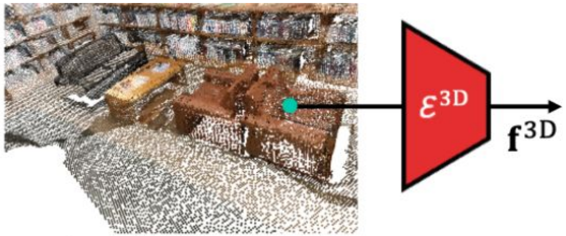
Introduction & Overview

Learning an encoder that takes in 3D Shapes to create per-pixel 3D embeddings

Problem: Available 3D datasets to small for good generalization

Ensembling datasets & Strategies for effective learning

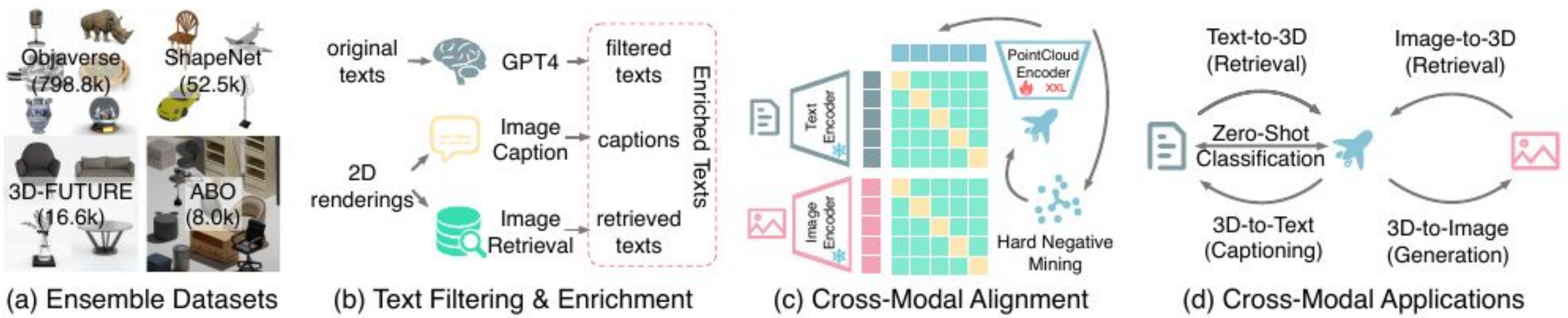
OpenScene



Input 3D Geometry

OpenShape: Scaling up 3D Shape Representation Towards Open-World Understanding

Introduction & Overview



OpenShape: Scaling up 3D Shape Representation Towards Open-World Understanding

Dataset ensembling

Four largest public 3D datasets ensembled (876k shapes)

ShapeNet, ABO, 3D-Future cover limited shapes and categories

Objaverse is more diverse but has uneven quality and distributions

OpenShape: Scaling up 3D Shape Representation Towards Open-World Understanding

Dataset ensembling: Objaverse

Objaverse is uploaded by web users, not human-verified for quality

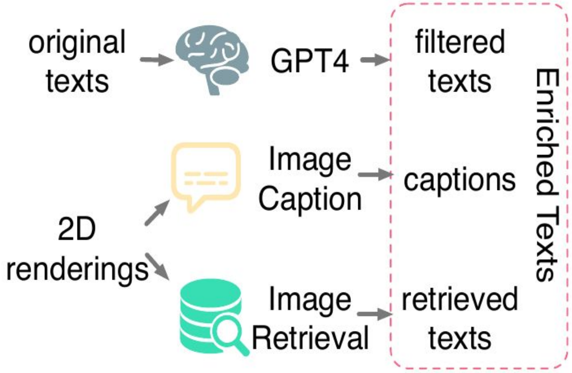
Text descriptions are noisy

Uninformative or inaccurate ground truth labels

OpenShape: Scaling up 3D Shape Representation Towards Open-World Understanding

Text Filtering and Enrichment

- GPT 4 filters out inaccurate or uninformative texts
- BLIP and Azure used to generate text descriptions from images
- k-NN images from LAION-5B retrieved using CLIP ViT-L index. Captions from these images also used.

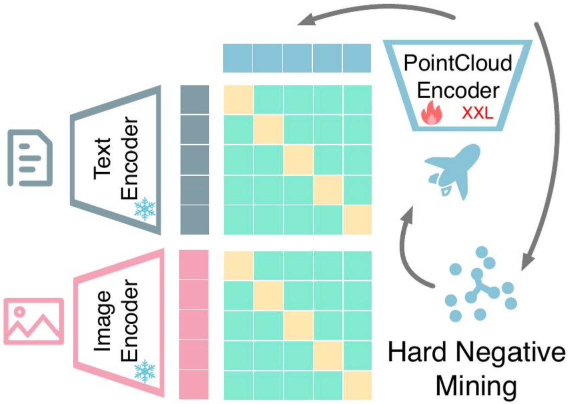


OpenShape: Scaling up 3D Shape Representation Towards Open-World Understanding
Training overview

Text and image encoders from CLIP are frozen

Sample of point cloud, image and text are encoded

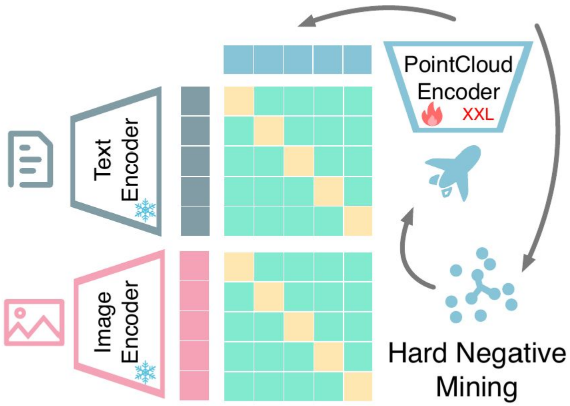
Trained to maximize matching pairs and minimize other embeddings



OpenShape: Scaling up 3D Shape Representation Towards Open-World Understanding

Hard Negative Mining

- Normal first round of training with random batches
- Second round, randomly select shapes and obtain k-NN neighbors of those shapes
- Increases likelihood of confusing pairs in a batch



Experiments & Results

OpenShape: Scaling up 3D Shape Representation Towards Open-World Understanding

OpenShape: Scaling up 3D Shape Representation Towards Open-World Understanding

Zero-Shot shape classification

OpenShape compared to existing zero-shot approaches

Method	training shape source	Objaverse-LVIS [12]			ModelNet40 [72]			ScanObjectNN [68]		
		Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
PointCLIP [82]	2D inferences, no 3D Training	1.9	4.1	5.8	19.3	28.6	34.8	10.5	20.8	30.6
PointCLIP v2 [84]	2D inferences, no 3D Training	4.7	9.5	12.9	63.6	77.9	85.0	42.2	63.3	74.5
ReCon [51]	ShapeNet	1.1	2.7	3.7	61.2	73.9	78.1	42.3	62.5	75.6
CG3D [19]	ShapeNet	5.0	9.5	11.6	48.7	60.7	66.5	42.5	57.3	60.8
CLIP2Point [24]	ShapeNet	2.7	5.8	7.9	49.5	71.3	81.2	25.5	44.6	59.4
ULIP-PointBERT (Official) [75]	ShapeNet	6.2	13.6	17.9	60.4	79.0	84.4	51.5	71.1	80.2
OpenShape-SparseConv	ShapeNet	11.6	21.8	27.1	72.9	87.2	93.0	52.7	72.7	83.6
OpenShape-PointBERT	ShapeNet	10.8	20.2	25.0	70.3	86.9	91.3	51.3	69.4	78.4
ULIP-PointBERT (Retrained)	Ensembled (no LVIS)	21.4	38.1	46.0	71.4	84.4	89.2	46.0	66.1	76.4
OpenShape-SparseConv	Ensembled (no LVIS)	37.0	58.4	66.9	82.6	95.0	97.5	54.9	76.8	87.0
OpenShape-PointBERT	Ensembled (no LVIS)	39.1	60.8	68.9	**85.3**	96.2	97.4	47.2	72.4	84.7
ULIP-PointBERT (Retrained)	Ensembled	26.8	44.8	52.6	75.1	88.1	93.2	51.6	72.5	82.3
OpenShape-SparseConv	Ensembled	43.4	64.8	72.4	83.4	95.6	97.8	**56.7**	78.9	88.6
OpenShape-PointBERT	Ensembled	**46.8**	**69.1**	**77.0**	84.4	**96.5**	**98.0**	52.2	**79.7**	**88.7**

OpenShape: Scaling up 3D Shape Representation Towards Open-World Understanding

Zero-Shot shape classification

OpenShape compared to existing zero-shot approaches

Method	training shape source	Objaverse-LVIS [12]			ModelNet40 [72]			ScanObjectNN [68]		
		Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
PointCLIP [82]	2D inferences, no 3D Training	1.9	4.1	5.8	19.3	28.6	34.8	10.5	20.8	30.6
PointCLIP v2 [84]	2D inferences, no 3D Training	4.7	9.5	12.9	63.6	77.9	85.0	42.2	63.3	74.5
ReCon [51]	ShapeNet	1.1	2.7	3.7	61.2	73.9	78.1	42.3	62.5	75.6
CG3D [19]	ShapeNet	5.0	9.5	11.6	48.7	60.7	66.5	42.5	57.3	60.8
CLIP2Point [24]	ShapeNet	2.7	5.8	7.9	49.5	71.3	81.2	25.5	44.6	59.4
ULIP-PointBERT (Official) [75]	ShapeNet	6.2	13.6	17.9	60.4	79.0	84.4	51.5	71.1	80.2
OpenShape-SparseConv	ShapeNet	11.6	21.8	27.1	72.9	87.2	93.0	52.7	72.7	83.6
OpenShape-PointBERT	ShapeNet	10.8	20.2	25.0	70.3	86.9	91.3	51.3	69.4	78.4
ULIP-PointBERT (Retrained)	Ensembled (no LVIS)	21.4	38.1	46.0	71.4	84.4	89.2	46.0	66.1	76.4
OpenShape-SparseConv	Ensembled (no LVIS)	37.0	58.4	66.9	82.6	95.0	97.5	54.9	76.8	87.0
OpenShape-PointBERT	Ensembled (no LVIS)	39.1	60.8	68.9	**85.3**	96.2	97.4	47.2	72.4	84.7
ULIP-PointBERT (Retrained)	Ensembled	26.8	44.8	52.6	75.1	88.1	93.2	51.6	72.5	82.3
OpenShape-SparseConv	Ensembled	43.4	64.8	72.4	83.4	95.6	97.8	**56.7**	78.9	88.6
OpenShape-PointBERT	Ensembled	**46.8**	**69.1**	**77.0**	84.4	**96.5**	**98.0**	52.2	**79.7**	**88.7**

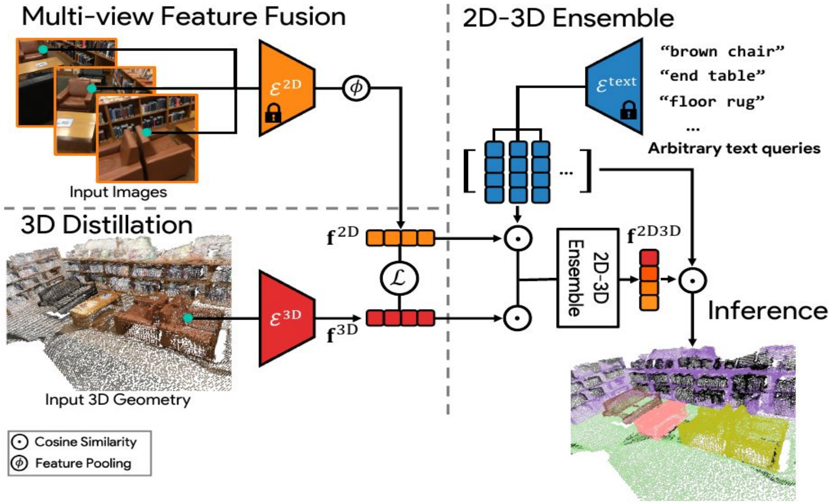
Future Work

Future Work

Improving OpenScene

Uses pixel-aligned features (ConceptFusion)

Distillation of 3D encoder (OpenShape)

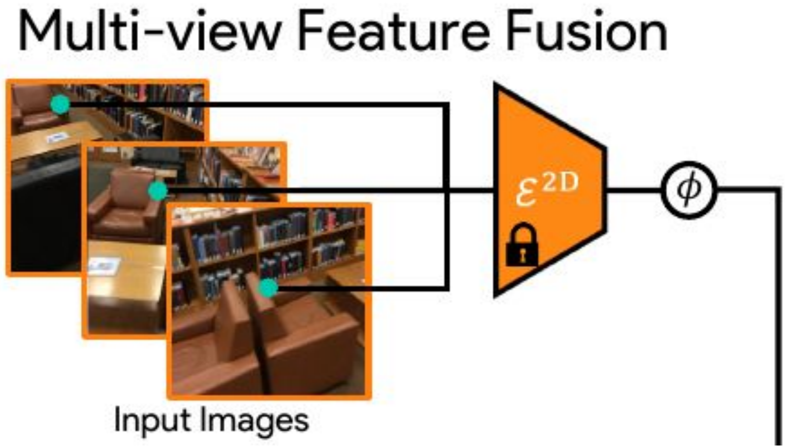


Future Work

Limitations of pixel-aligned features

Embeddings depend on image viewpoint

Few corresponding images per point



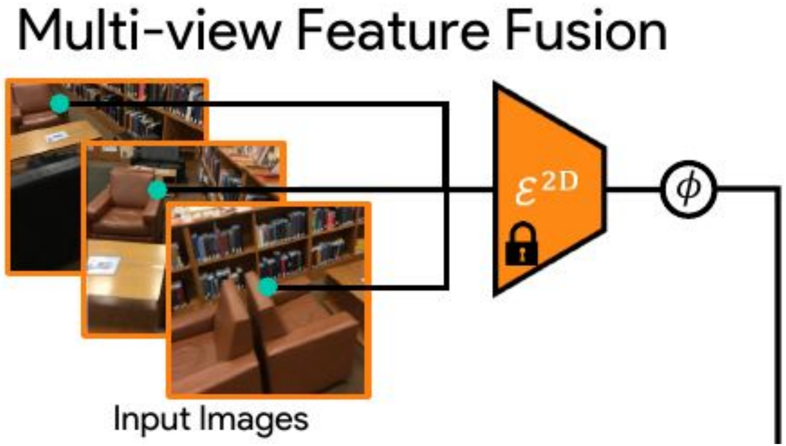
Future Work

Idea: Apply generative AI

1. Identify objects in view
2. Generate more images of the objects using the point cloud

Apply NERF to generate different viewing angles

Take pictures of objects inside the point cloud then improve quality

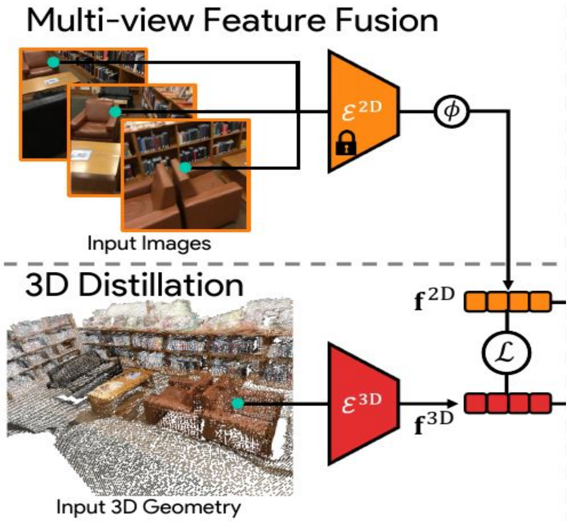


Future Work

Limitation of OpenScene 3D Distillation

Training relies on images of the specific scene

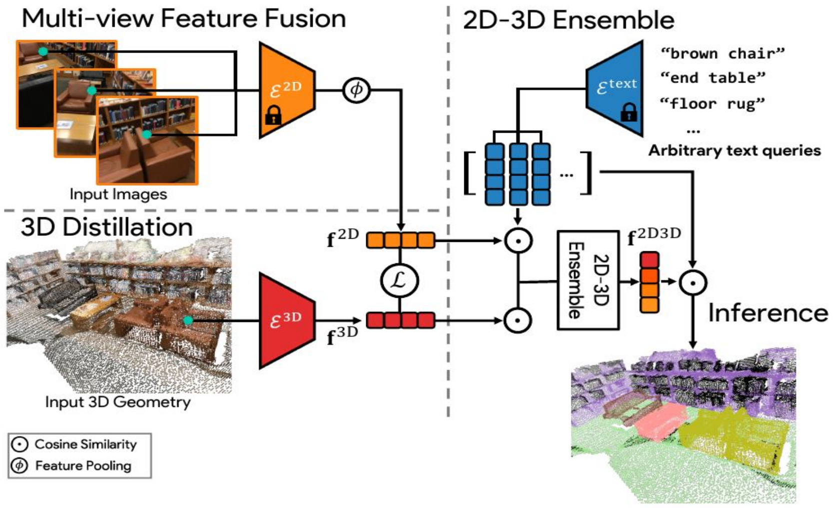
Solution: Scale up 3D encoder training to be more generalizable (e.g OpenShape)



Future Work

Better performance during inference

- More accurate generation of f^{2D} through synthetic image generation
- More accurate generation of f^{3D} through scaling up ε^{3D} training
- More accurate features after ensemble leading to more accurate and flexible inference



Summary

Open Vocabulary 3D Scene Understanding
Summary

- Modern approaches generate embeddings as the semantic anchor between 3D points, images, and queries
- CLIP is a popular and proven candidate for generating these embeddings
- Embeddings are be generated using two methods:
    - Using 2D images of the scene to extract embeddings of each pixel using CLIP
    - Training a 3D encoder which generates embeddings in the same space as CLIP embeddings
- Work improving the accuracy and generalization of these two methods will increase performance of 3D scene understanding in the future

Thank you!

Question & Answer