AT3DCV - Group 1

Multimodal 3D Understanding

Aleksandar Jevtic
Orhun Aksoy
Idil Ünlü

Technische Universität München

Content

- Motivation
- Related Work
- Our Work
- Experiments and Results
- Demo
- Conclusion

Motivation

Why Multimodal Understanding?

Example:
Service robot for care and security

Source: https://www.cnet.com/home/this-robot-isnt-going-to-replace-your-in-home-nurse-yet/

Example:
Service robot for care and security

Source: https://www.cnet.com/home/this-robot-isnt-going-to-replace-your-in-home-nurse-yet/

Needs understanding of its environment

Source: https://www.cnet.com/home/this-robot-isnt-going-to-replace-your-in-home-nurse-yet/

Needs understanding of its environment

Source: https://www.cnet.com/home/this-robot-isnt-going-to-replace-your-in-home-nurse-yet/

but…

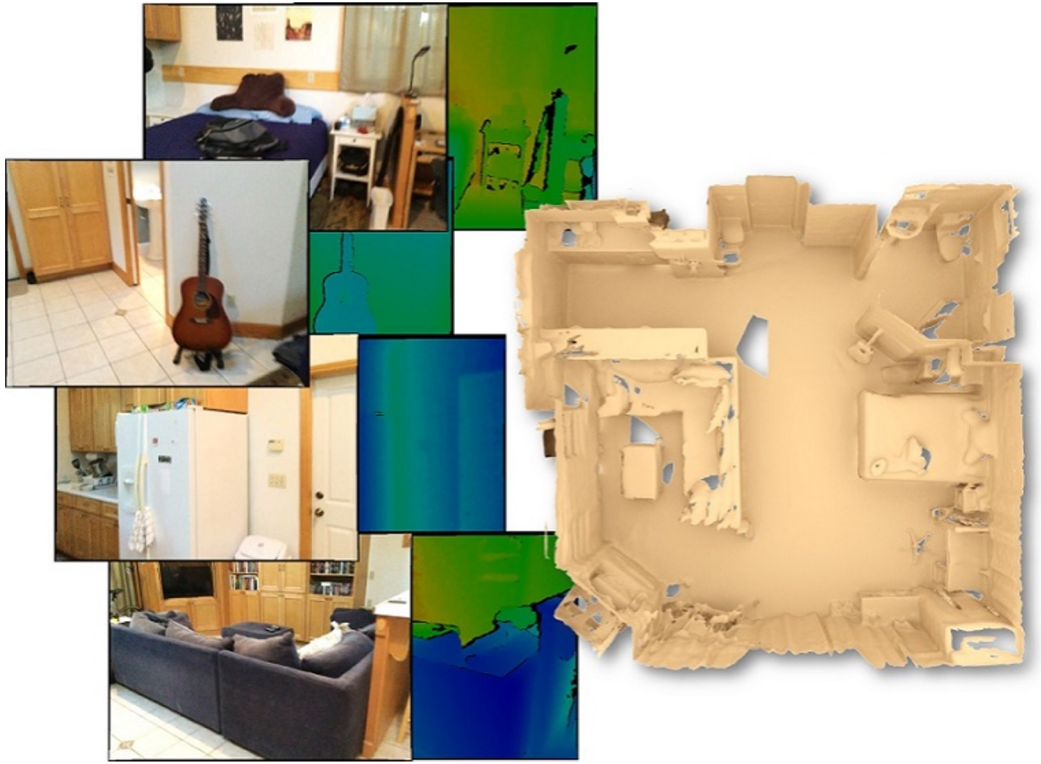
Real world can be very challenging!

Source: Dai, Angela and Chang, Angel X. and Savva, Manolis and Halber, Maciej and Funkhouser, Thomas and Nießner, Matthias, “ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes.” CVPR. 2017.

Use all available information:

- 2D images
- 3D data
- Natural language

→ Multimodal Understanding



Source: Dai, Angela and Chang, Angel X. and Savva, Manolis and Halber, Maciej and Funkhouser, Thomas and Nießner, Matthias, “ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes.” CVPR. 2017.

Research Questions

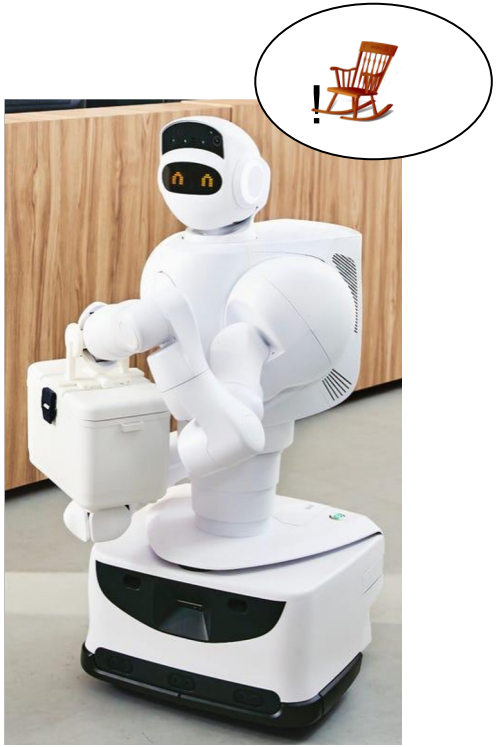
Can we detect novel and **uncommon objects** in 3D scenes?

Capabilities of **State of the Art**?

- Open-set
- Zero-shot

How can we **improve** it?

Idea: Employ visual descriptors by LLMs

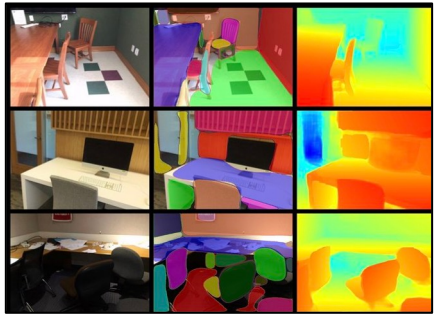


Source: https://www.cnet.com/home/this-robot-isnt-going-to-replace-your-in-home-nurse-yet/

Our Solution

- Extend existing datasets with uncommon objects
  - To test the limits of the CLIP and OpenScene
- Integrating LLMs into the pipeline
  - To test if using descriptors would improve the model
- Aggregate the descriptors' results and evaluate

ScanNet - 3D Scene



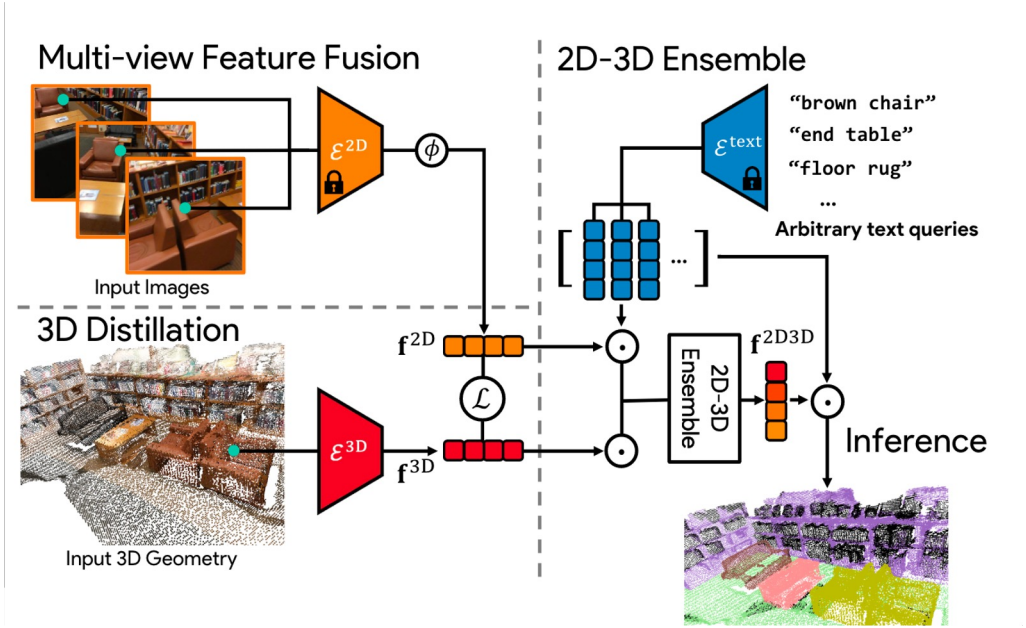
\+ Augmentations

Source: https://www.cnet.com/home/this-robot-isnt-going-to-replace-your-in-home-nurse-yet/

Related Work

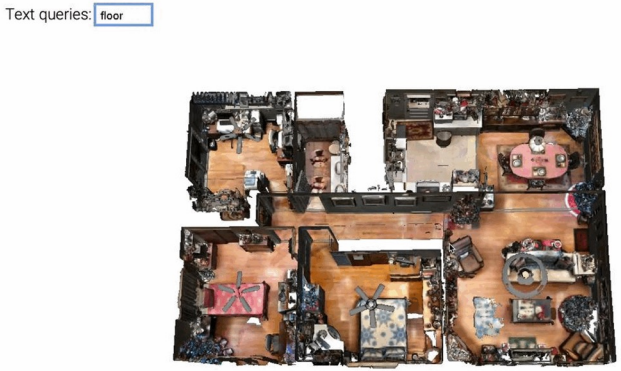
OpenScene

- 2D or 3D features ensembled with CLIP features
- Co-embeds dense 3D features with image pixels and text
- CLIP features used for queries at the end



OpenScene

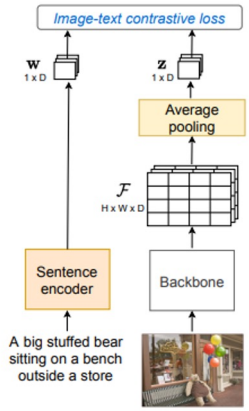
- Zero-shot
- Open vocabulary queries (objects, materials, activities, …)
- Heat map for text query



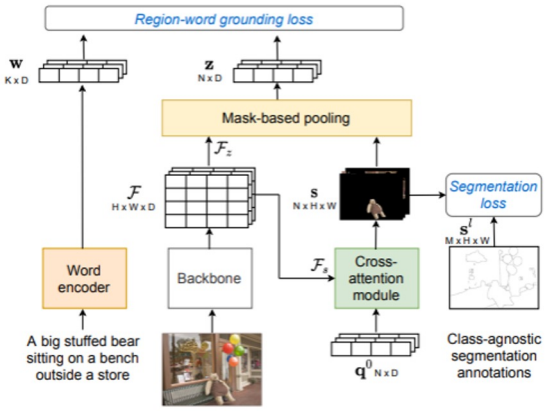
Source: S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, "Openscene: 3d scene understanding with open vocabularies," CVPR, 2023.

OpenSeg

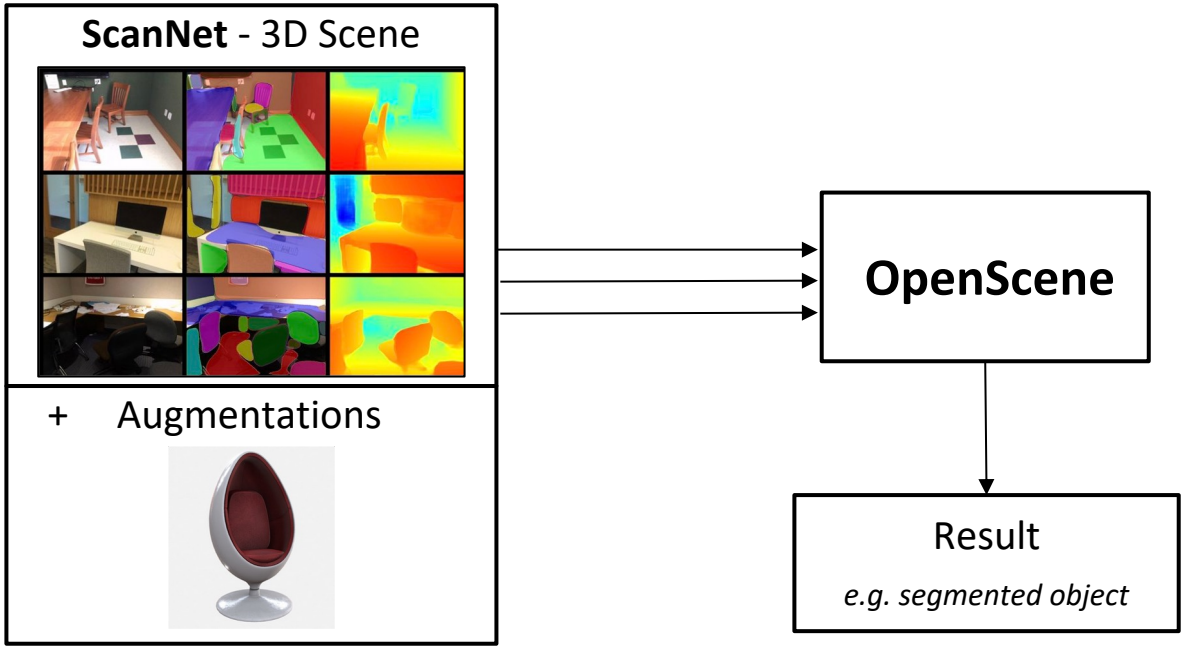
- Semantic Segmentation
- Zero-Shot
- Based on pre-trained CLIP
- Pre-defined masks for objects



(a) ALIGN / CLIP



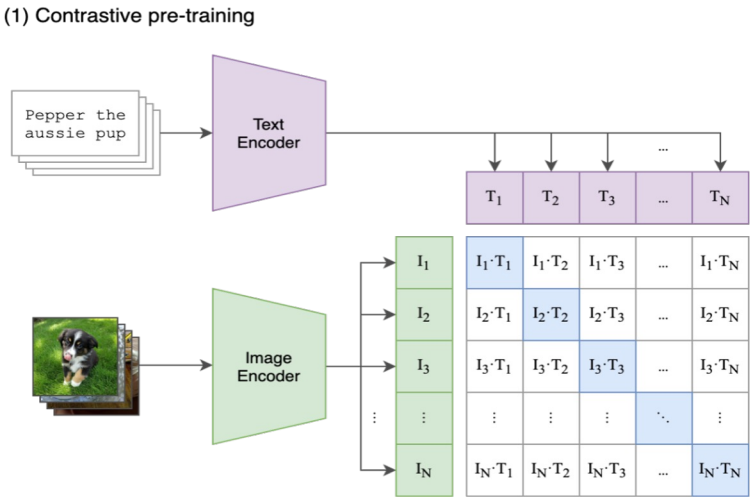
(c) OpenSeg (ours)

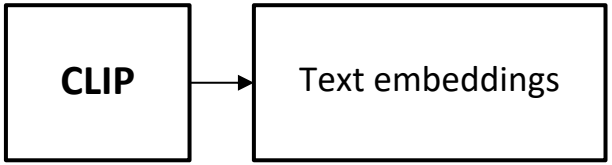


Source: G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, “Scaling open-vocabulary image segmentation with image-level labels,” ECCV 2022

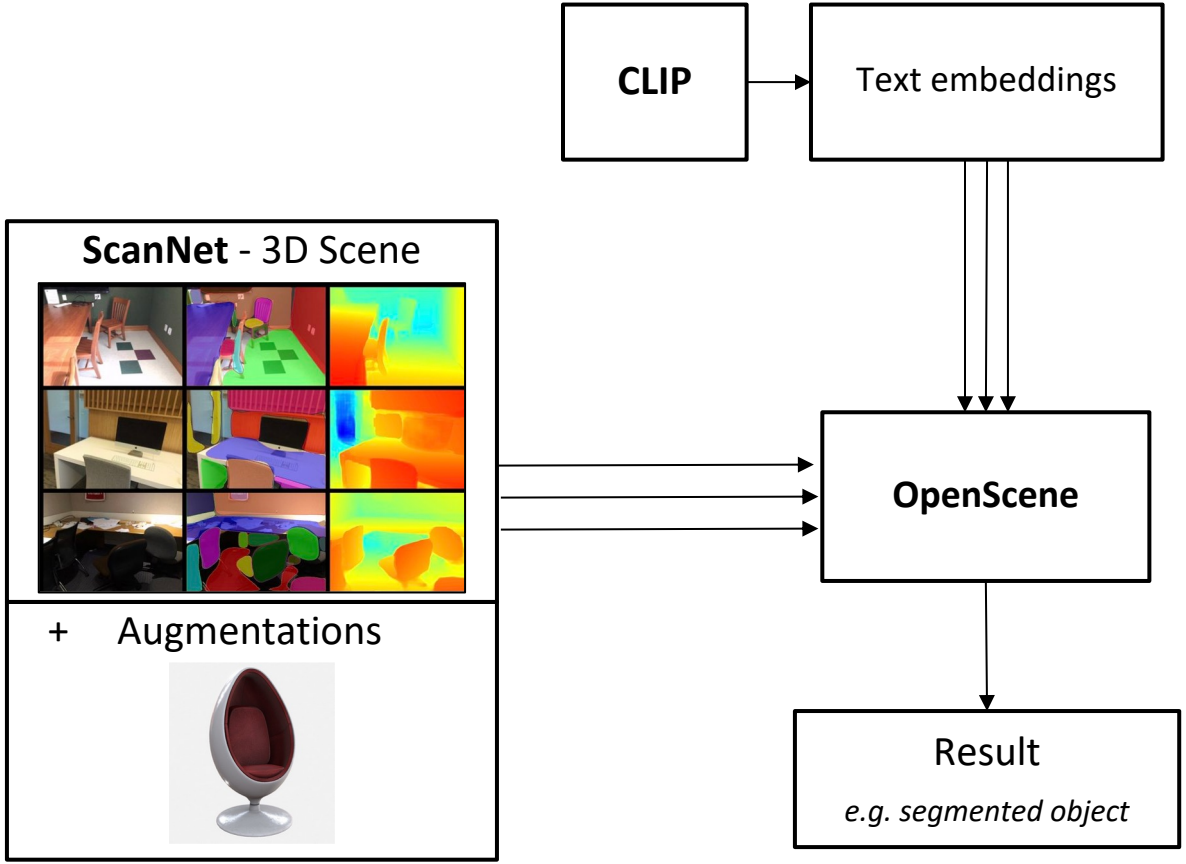
Contrastive Language-Image Pre-Training

- Zero-shot
- Combined text and image embeddings into a single space



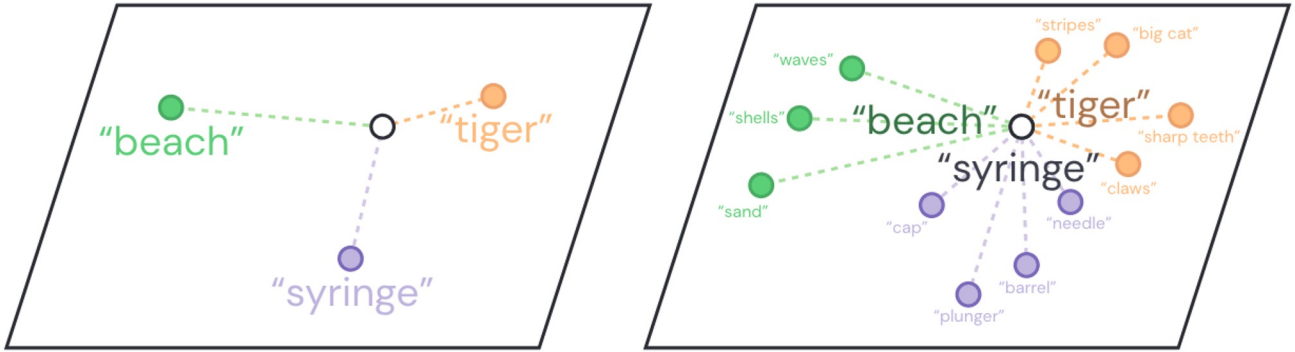


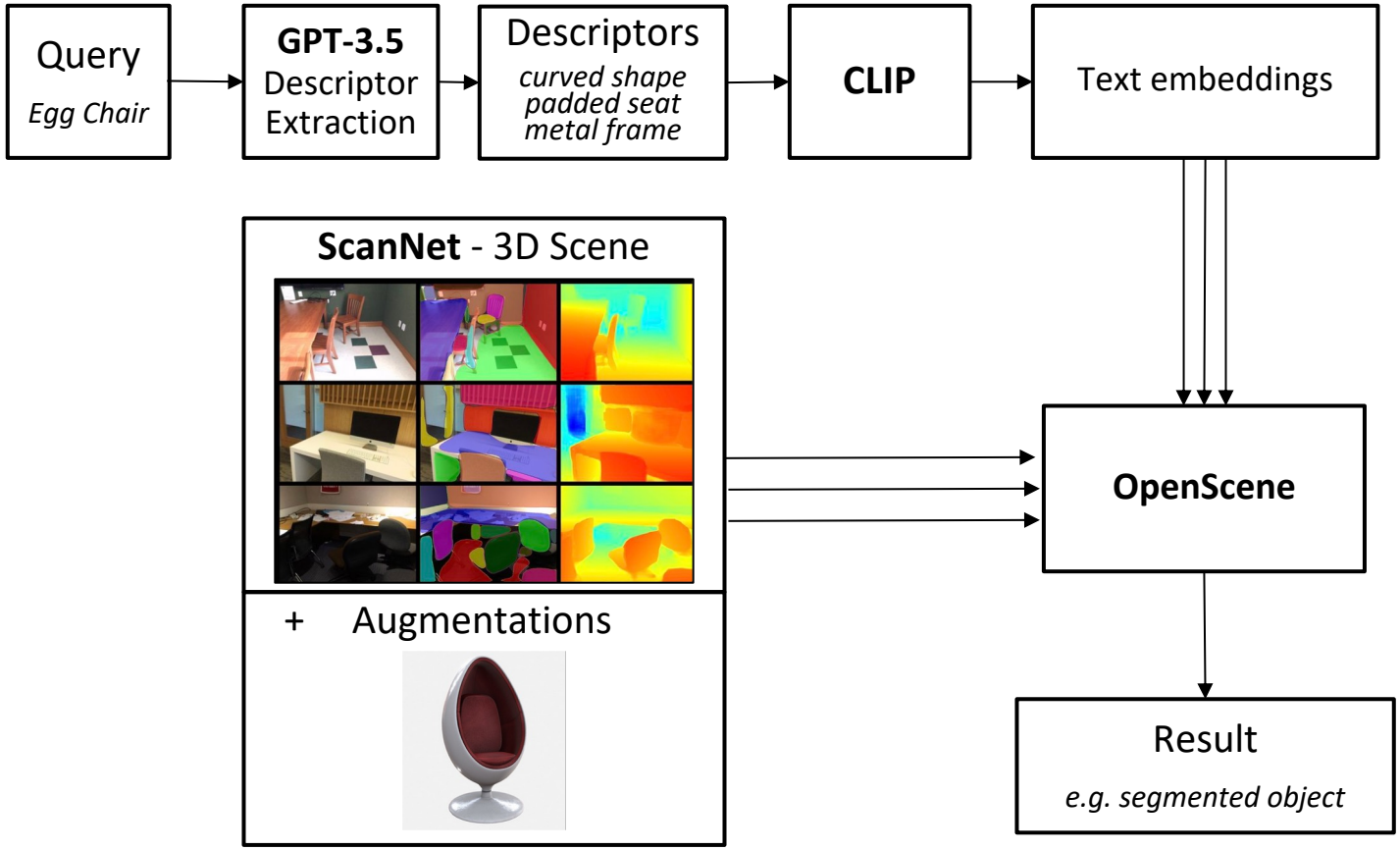
Source: A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," ICML, 2021.


CLIP
Text embeddings
ScanNet - 3D Scene
+ Augmentations
OpenScene
Result
e.g. segmented object

Visual Classification via LLM Descriptions

- Language as an internal representation for visual recognition
- VLM checks for features instead of just the class name
- Detailed prompts increase IoU

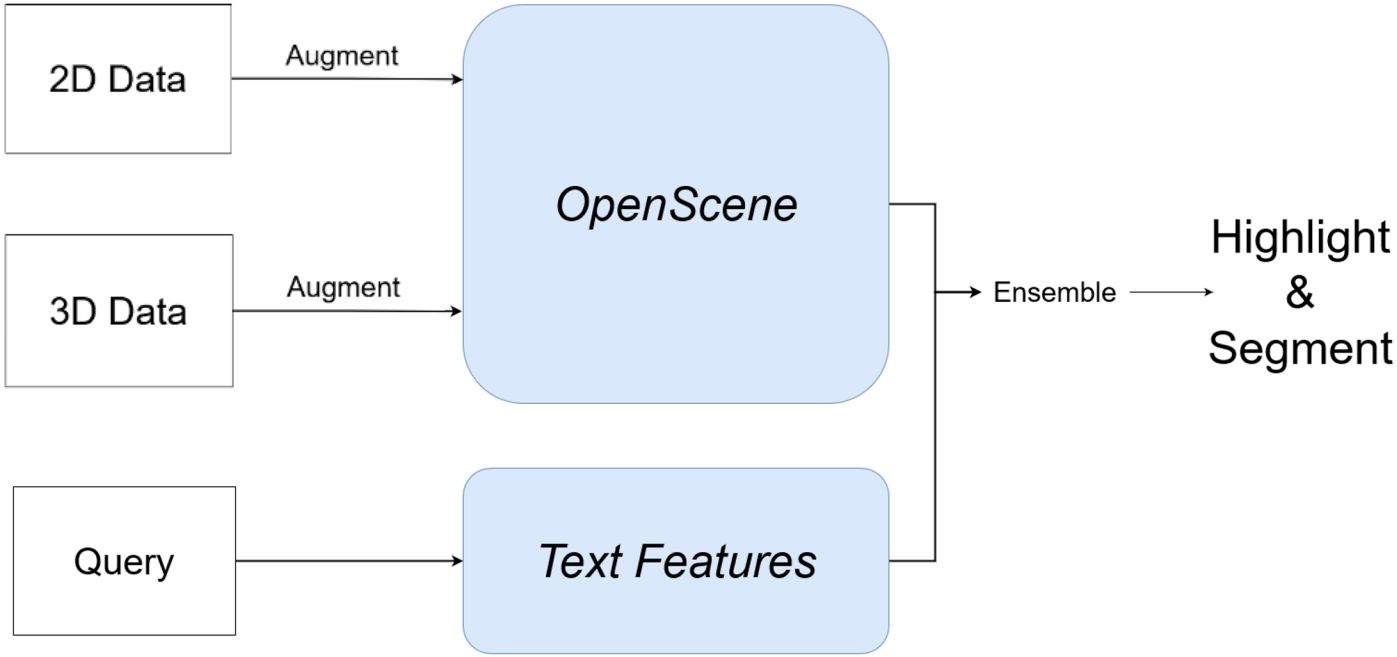



Query
Egg Chair
GPT-3.5
Descriptor
Extraction
Descriptors
curved shape
padded seat
metal frame
CLIP
Text embeddings
ScanNet - 3D Scene
+ Augmentations
OpenScene
Result
e.g. segmented object

Our Work

Technische Universität München

Pipeline Overview



Source: S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, “Openscene: 3d scene understanding with open vocabularies,” CVPR, 2023.

Augmenting Data with Novel Objects

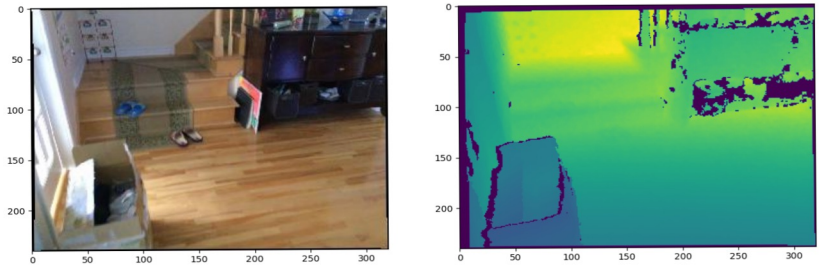
Raw 3D Scene

Augmented 3D Scene

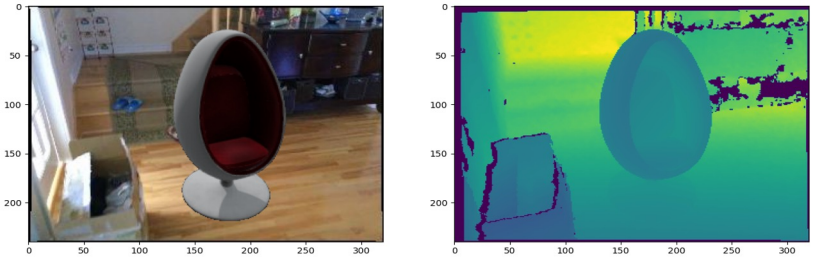
- Adding 3D models to 3D point cloud

Source: S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, “Openscene: 3d scene understanding with open vocabularies,” CVPR, 2023.

Augmenting Data with Novel Objects



Raw Multi-view Images

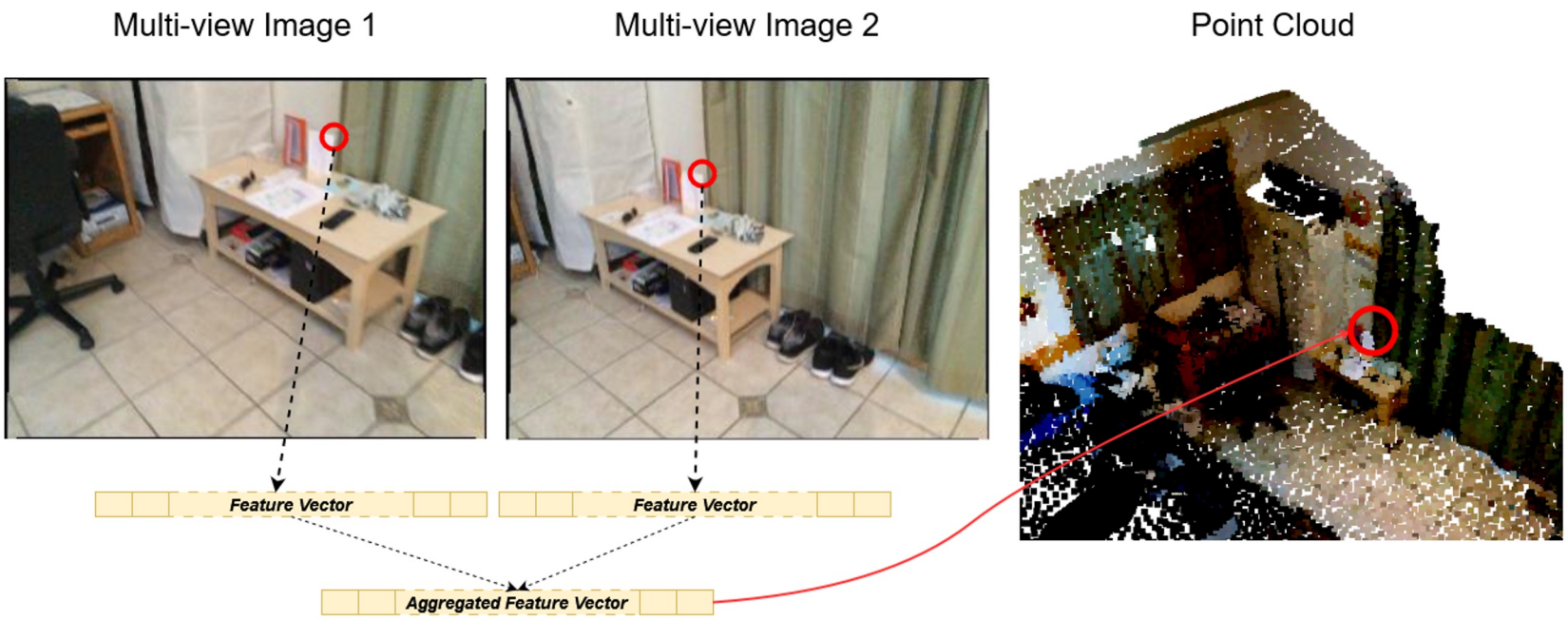


Augmented Multi-view Images

- Projecting 3D model to 2D images

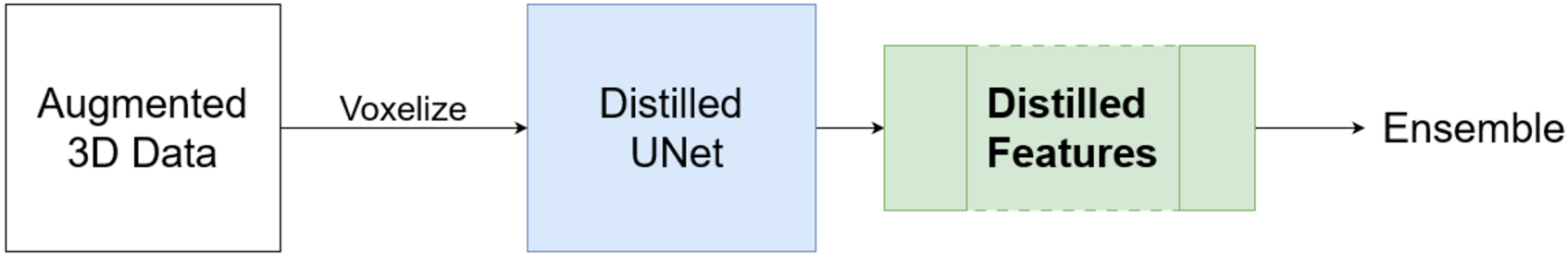
Source: S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, “Openscene: 3d scene understanding with open vocabularies,” CVPR, 2023.

2D Features & Fusion



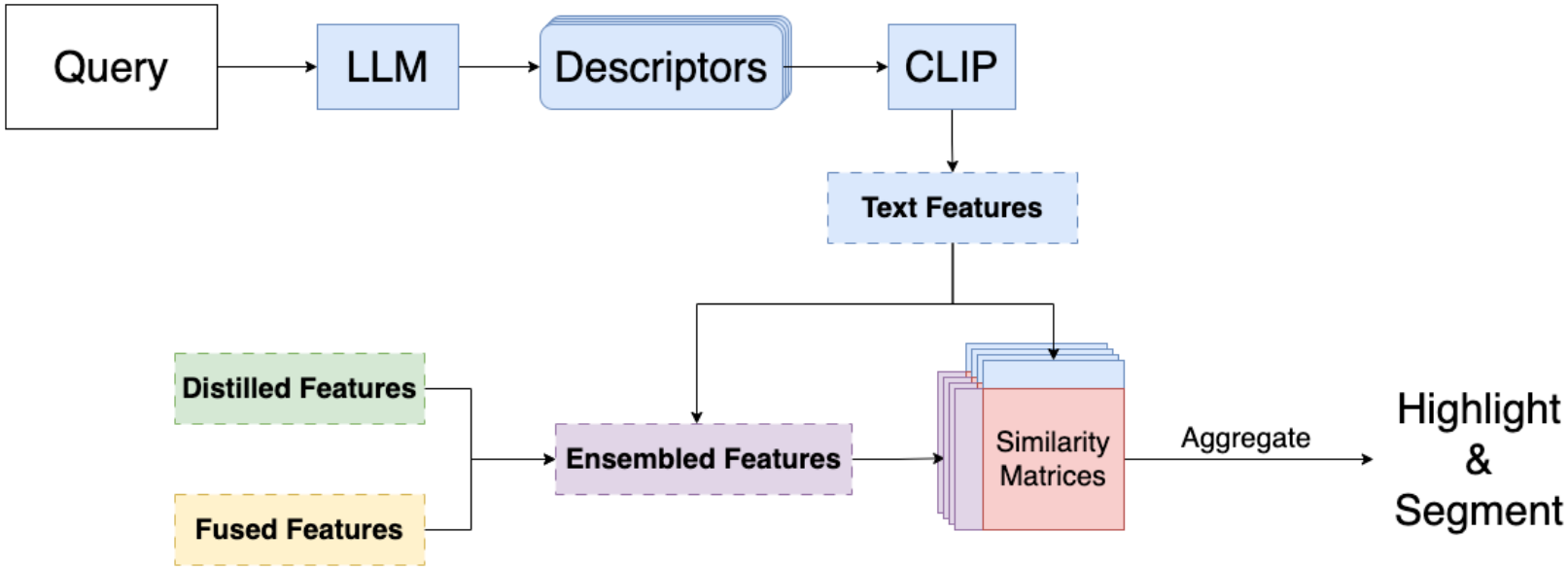
Source: S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, “Openscene: 3d scene understanding with open vocabularies,” CVPR, 2023.

3D Features



Source: S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, “Openscene: 3d scene understanding with open vocabularies,” CVPR, 2023.

Ensemble and Text Features



Source: S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, "Openscene: 3d scene understanding with open vocabularies," CVPR, 2023.

Experiments/Results

Experiment on Bird Species

Nicobar Pigeon

Eastern Rosella

Source: *Nicobar pigeon* (2023) *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Nicobar_pigeon.,
Eastern rosella (2023) *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Eastern_rosella.

Experiment on Bird Species

Descriptors from LLM:

Nicobar Pigeon

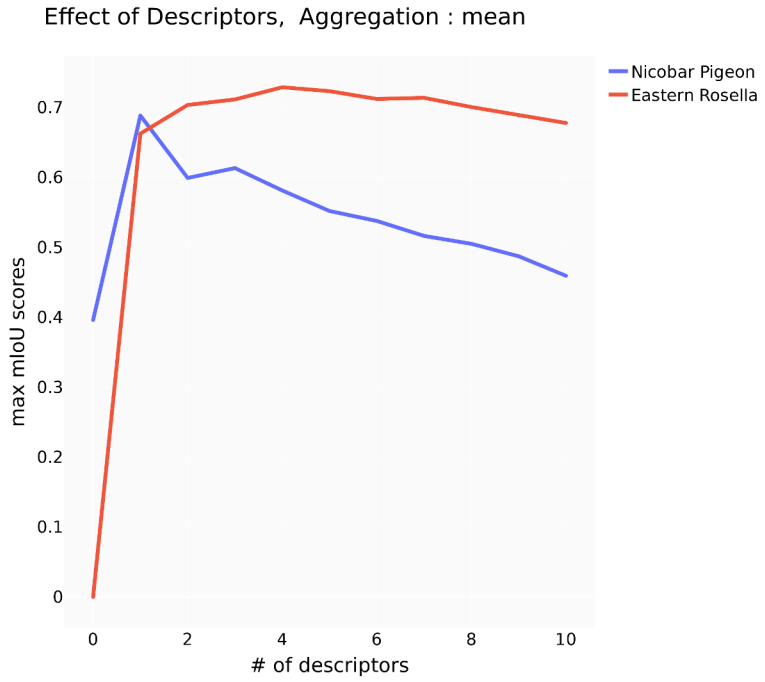
'Green-blue plumage', 'Metallic-sheen feathers',
'Slender body', 'Long tail feathers',
'Light brown head', 'Red beak',
'White eye-ring', 'Pink feet',
'Dark neck patch', 'Yellow shoulder stripe'

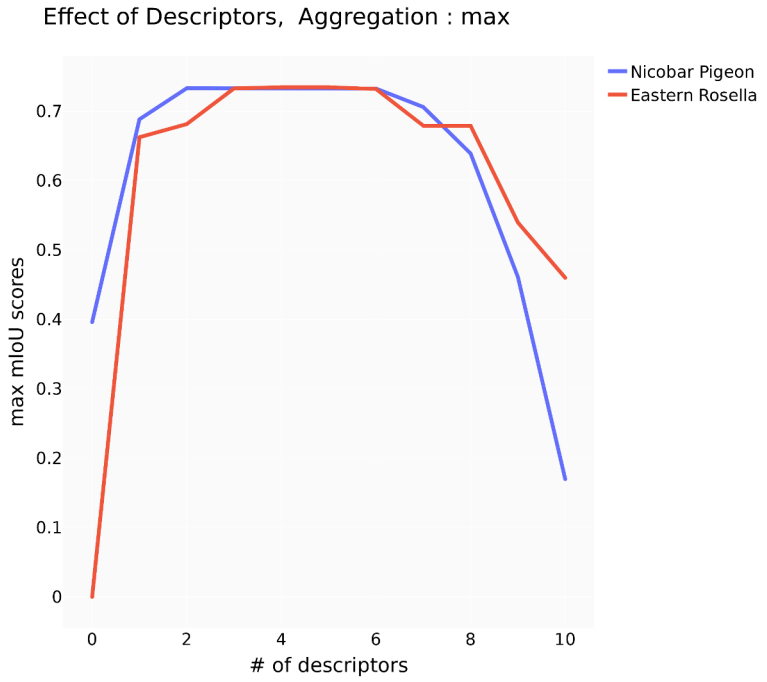
Eastern Rosella

'Red head', 'Red shoulder patches'
'Blue wings', 'White breast'
'Yellow belly', 'Orange-red tail'
'Black beak', 'Blue-green back'
'White eye-ring', 'Long tail feathers'

Prompt: ‘a bird which is/has <feature>’

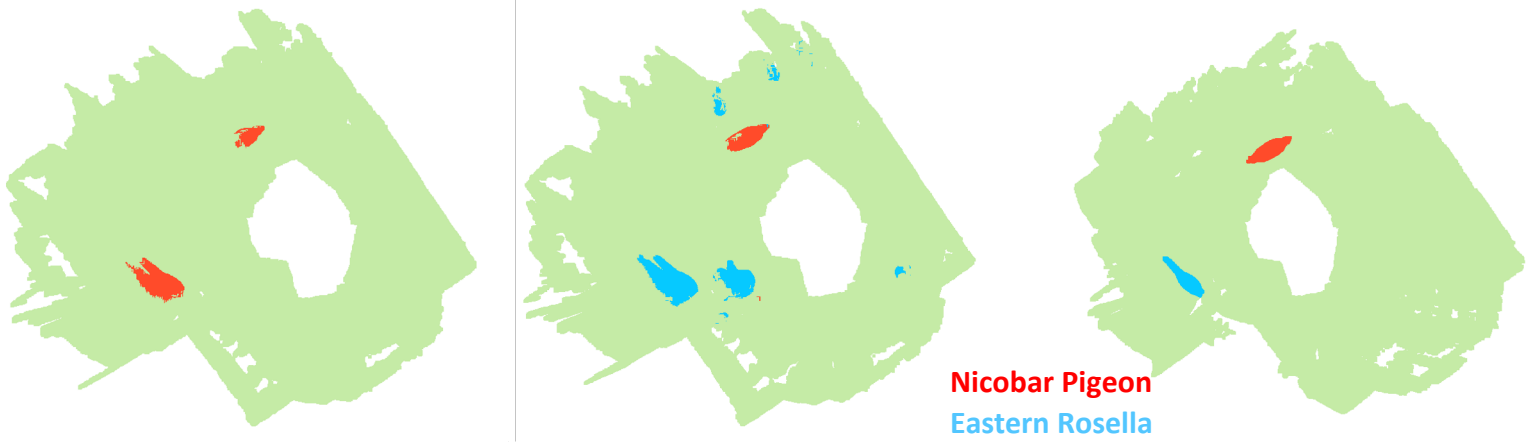
Experiment on Bird Species





Experiment on Bird Species

query by name | *query by descriptors* | *ground truth*



- Nicobar Pigeon, descriptors for highest IoU:

'Metallic-sheen feathers', 'Slender body', 'Long tail feathers', 'Light brown head', 'Red beak', 'Dark neck patch'

- Eastern Rosella, descriptors for highest IoU:

'Red shoulder patches', 'Yellow belly', 'Orange-red tail', 'Black beak', 'White eye-ring', 'Long tail feathers'

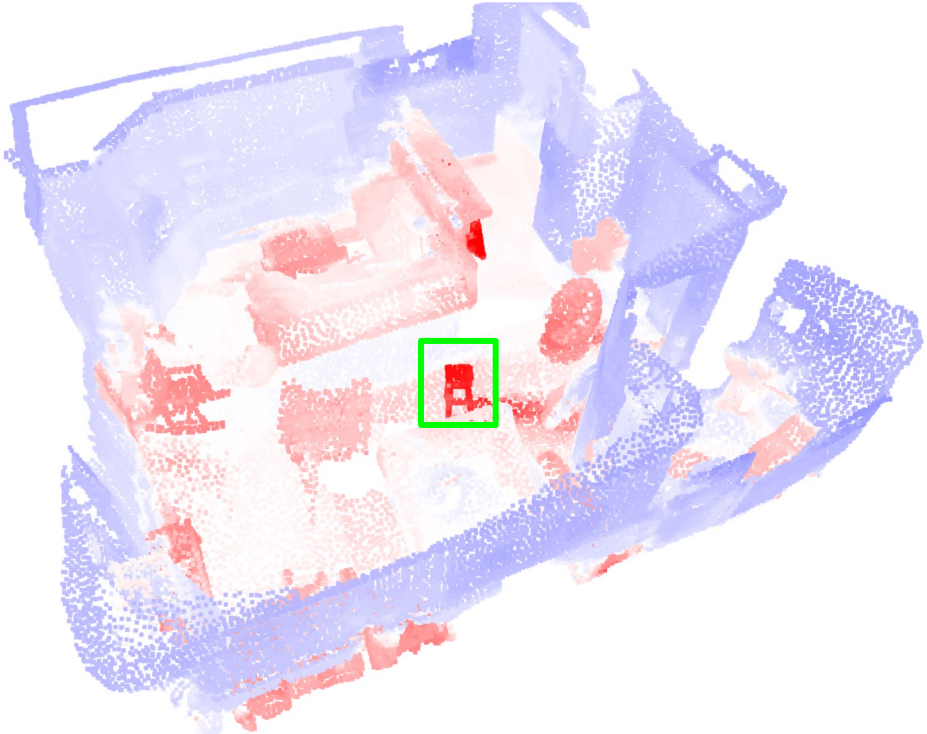
Experiment on Chairs

Augmented ScanNet scene with different chair models

- Stool
- Armchair
- Ball chair
- Rocking chair

Query:

stool



Descriptors for *stool*

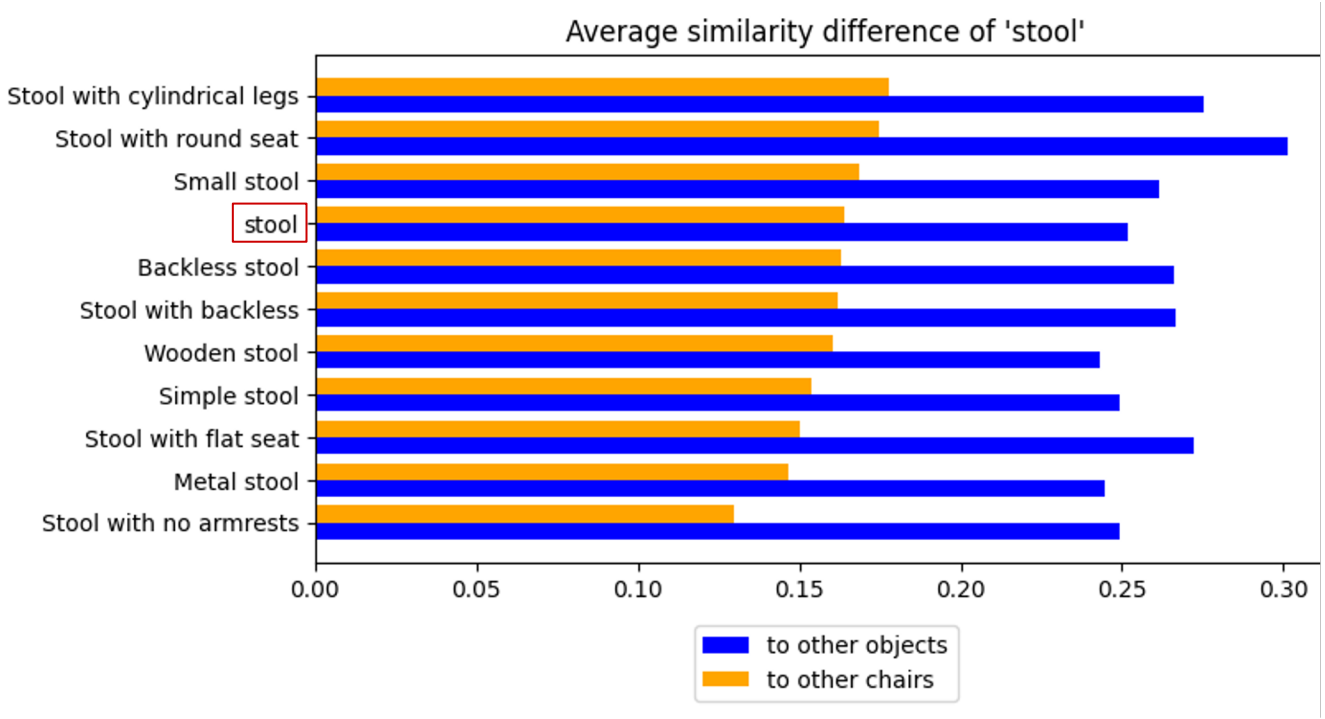
Regular query

- CLIP model already very accurate

Queries with visual descriptors

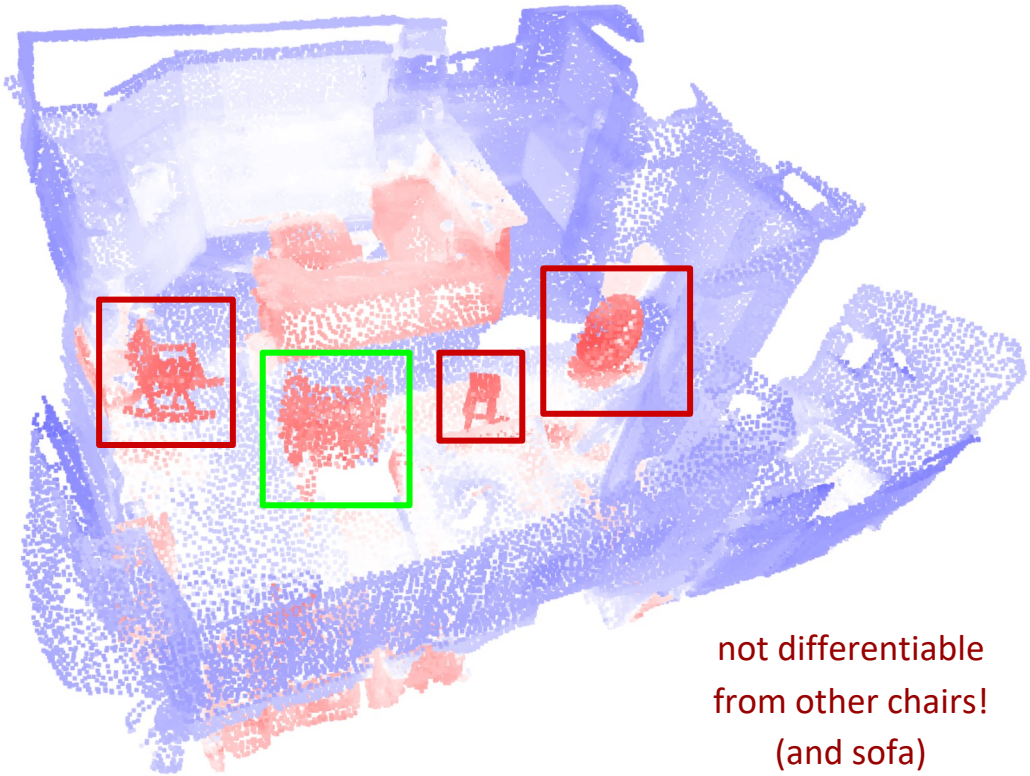
- Perform slightly worse on average
- some do not apply (*metal stool*)
- however: Reflect different aspects

Descriptors for *stool*



Query:

armchair



not differentiable
from other chairs!
(and sofa)

Descriptors for *armchair*

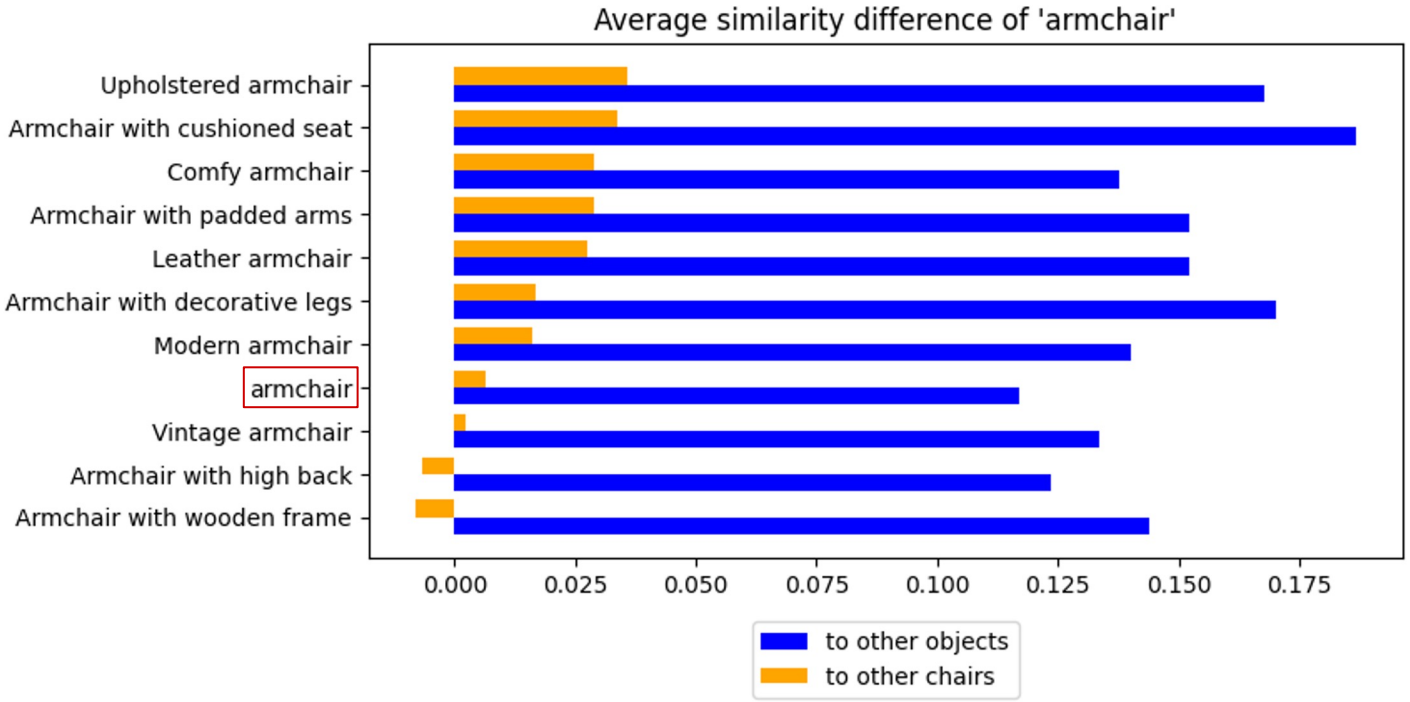
Regular query

- CLIP model has problems with concept
- Difference to other chairs is not captured

Queries with visual descriptors

- Perform better on average
- again: some do not apply (*armchair with wooden frame*)

Descriptors for *armchair*



Aggregation of descriptors

Average similarity

Quantile / Minimum ↔ Median ↔ Maximum

- # of descriptors that reflect object
- for this experiment: Median (0.5)

(Aggregation in learned weights, space)

Demo

Classification of Chairs

jupyter chairs_live_demo Last Checkpoint: 20 hours ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

```
    pc_colors_gt = darken_color_factor * filtered_pc_c
    for cl in range(0, len(colors)):
        pc_colors_gt[filtered_pc_labels == cl] = colors[cl]

    pcd_class = o3d.geometry.PointCloud()
    pcd_class.points = o3d.utility.Vector3dVector(filtered_pc)
    pcd_class.colors = o3d.utility.Vector3dVector(pc_colors_classification)

    gt_shift = [-7.5,-2.5,0]
    pcd_gt = o3d.geometry.PointCloud()
    pcd_gt.points = o3d.utility.Vector3dVector(filtered_pc + gt_shift)
    pcd_gt.colors = o3d.utility.Vector3dVector(pc_colors_gt)

    o3d.visualization.draw_geometries([pcd_class, pcd_gt])

button_class.on_click(visualize_classification)
```

Descriptors

none | combined | bare

Features

ensembled | fused | distilled

✔ Visualize

In [ ]:

Conclusion

Findings

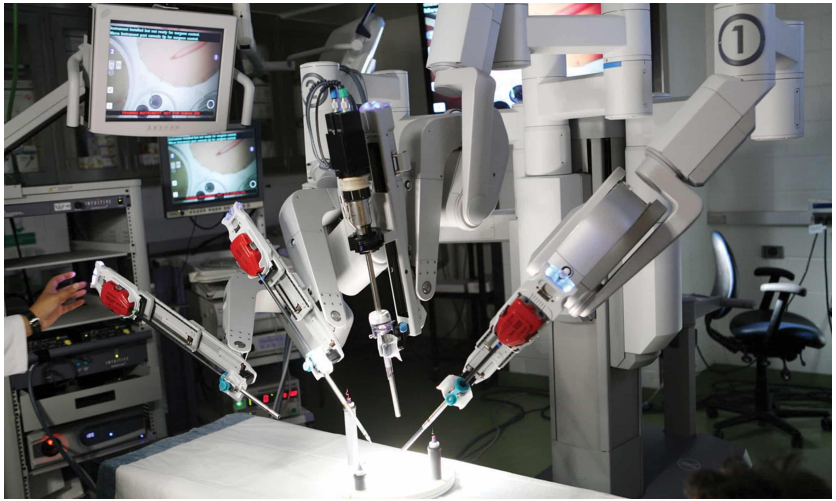
- Choice of descriptors is critical
  - Good number: 4-6
  - Conciseness

- Types of features
  - Good results with 2D features!
  - 3D features did not generalize well

- Query format

Future Work

- Improve Augmentation
  - Lighting, Anti-aliasing
- Better descriptors
  - Other LLMs, flexible
- Type of Aggregation
  - Learned weights, use priors, space
- Other segmentation models (LSeg)



Many possibilities!

Sources: https://www.automationalley.com/articles/industrial-robots-push-into-new-applications-and-industries
https://chicagohealthonline.com/robots-in-the-o-r/

Thanks for listening!

? ? ?