Final Project Proposal

EE367 / CS448I: Computational Imaging | Winter 2025
Codey Sun, Yin-Li Liu

Motivation

Open-vocabulary semantic segmentation describes the ability to extract the most relevant parts of a given image with proper text prompts. This capability is necessary for many applications spanning robotics, AR/VR, and general computer vision tasks. Current methods rely on large pretrained models like Segment Anything (SAM) and CLIP to extract these semantic embeddings, while these methods struggle when querying fine-grained parts of a scene. This project aims to improve open-vocabulary segmentation on the part level, allowing users to perform precise fine-grain segmentations from language prompts.

Related work

The Segment Anything Model (SAM) is a large pretrained model to extract edge-based segments of a 2D image [1]. Contrastive Learning Image Pretraining (CLIP) is a model that correlates images with text, allowing for open-vocabulary applications [2].

Together, work such as LERF [3] has lifted these 2D segments and labels to 3D by embedding them into a neural radiance field [4]. Given any text prompt, one can then compare the cosine similarities of the CLIP embeddings to extract the relevant 3D segments. Ultrametric feature fields are a similar 3D segmentation method that aggregates multi-view SAM segmentations using an ultrametric-based contrastive learning objective, effectively organizing segments into a hierarchical tree [5]. This method does not embed language, but its hierarchical tree structure may be useful for chain-of-thought prompting [6], which enables advanced reasoning in large language models.

Project goal

The inverse problem we are trying to solve is extracting part-level semantic information from a scene given labels from an image-to-text model. We can describe the inverse problem as the following:

$$y = Ax$$
y: $\mathit{image\ and\ segmentation\ with\ semantic\ labels}$
A: $\mathit{image\ segmenter}\ +\ \mathit{Chain} - \mathit{of} - \mathit{Thought\ reasoning\ model\ for\ segments\ labeling}$
x: $\mathit{semantic\ of\ the\ scene}$

Current methods achieve this by cropping masked segments (retrieved via SAM) and encoding those segments with a CLIP model; semantics are subsequently retrieved by interpreting the CLIP embeddings. However, this measurement model is noisy due to the limitations of CLIP to encode fine-grained parts. The left image in Fig. 1 shows an example input of a fine-grain part into CLIP. This signal is difficult to interpret by CLIP without the context of the whole chair. More specifically, the model will not know this is the central support of a chair without the “prior” knowledge of the object.

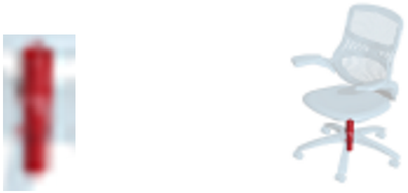

Figure 1. Cropped segment of a chair’s central support (left) vs. chair’s central support with context (right)

We aim to improve the measurement model by feeding the image encoder not only the cropped segment but also the context of what this part belongs to, i.e., a prior, as shown in the right image of Fig. 1. We will use the Chain-of-Thought reasoning capabilities of large vision-language models to chain the context of the object’s part hierarchy in the prompt. We imagine this will improve the output language embeddings compared to the traditional method, which only feeds a cropped image into the vision-language model. In turn, better semantics can be extracted from the scene.

Timeline and milestones

- Week 1: From input multiview images, segment the images and construct a part tree hierarchy
- Week 2: Create AI agent to interpret cropped image segments with context
- Week 3: Evaluate the method against traditional CLIP methods using the PartNet dataset

Reference

[1] Kirillov, Alexander, et al. "Segment anything." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[2] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

[3] Kerr, Justin, et al. "Lerf: Language embedded radiance fields." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[4] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM* 65.1 (2021): 99-106.

[5] He, Haodi, et al. "View-Consistent Hierarchical 3D Segmentation Using Ultrametric Feature Fields." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.

[6] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.