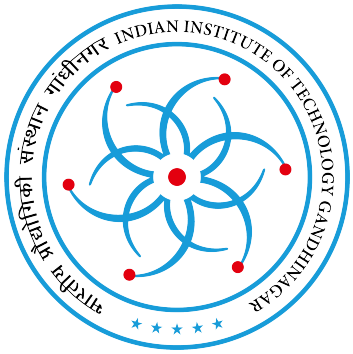
Indian Institute of Technology Gandhinagar
Computer Vision, Imaging, and Graphics (CVIG) Lab



CS 399 Mid-Sem Report
INDIAN INSTITUTE OF TECHNOLOGY GANDHINAGAR
Palaj, Gandhinagar - 382355

Open-vocabulary Image Segmentation

Submitted by

Guntas Singh Saran
Computer Science and Engineering (22110089)

Under the guidance of

Prof. Shanmuganathan Raman
Jibaben Patel Chair in Artificial Intelligence and Associate Professor
Electrical Engineering & Computer Science and Engineering, IIT Gandhinagar

Prajwal Singh
Ph.D. Scholar
Computer Science and Engineering, IIT Gandhinagar

Contents

Chapter 1

Research Topic

The finalized research topic is **Open-vocabulary Image Segmentation**. We are targeting the paper titled *Hierarchical Open-vocabulary Universal Image Segmentation (HIPIE)* [1]. This paper focuses on segmenting images based on arbitrary text descriptions, while accounting for inherent segmentation ambiguity due to different levels of granularity. The approach introduces a hierarchical representation that captures different semantic levels within the learning process, while existing methods typically sidestep this challenge. The HIPIE model employs both a

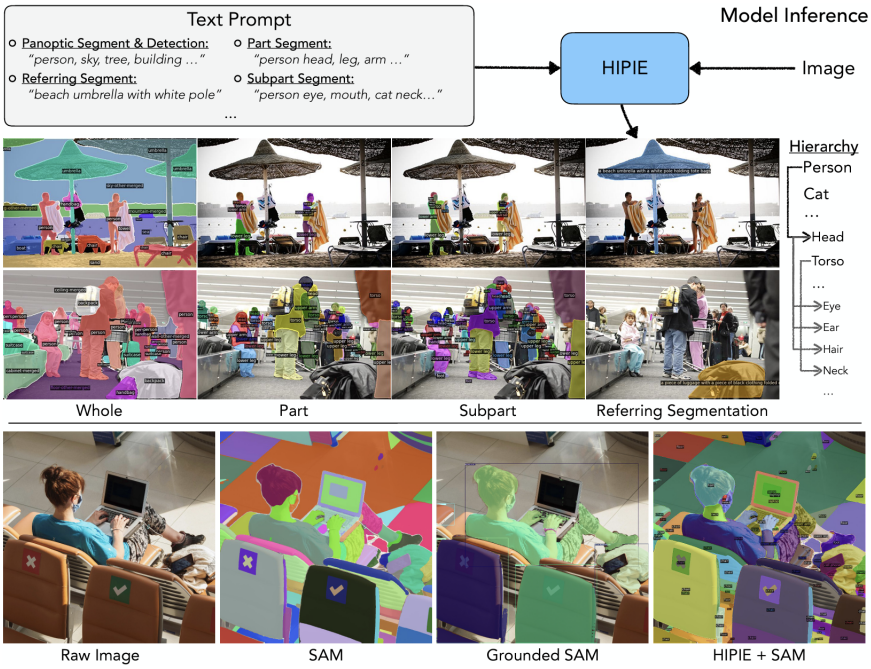


Figure 1.1: HIPIE, given an image and a set of arbitrary text descriptions, provides hierarchical semantic, instance, part, and subpart-levelimage segmentations. Adapted From [1].

Stuff Decoder and a **Thing Decoder** to tackle open-vocabulary segmentation, and achieves state-of-the-art results across over 40 benchmark datasets, including ADE20K, COCO, Pascal-VOC Part, and others. The model performs at multiple levels of comprehension, such as semantic, instance, and part-level segmentation.

Chapter 2

Objectives

2.1 Main Objective

The primary objectives of our research are:

1. **Extending HIPIE's Vocabulary**: We aim to add annotations for newer Bag of Words (BoW) that are not included in HIPIE's current vocabulary. This may involve using NLP-based techniques, such as identifying synonyms for existing words in the model's vocabulary, and expanding the model's understanding of textual prompts.

2. **Adding Adapters to Decoders**: In HIPIE's architecture, we plan to insert **Adapters** into both the Stuff Decoder and Thing Decoder to explore their impact on the segmentation accuracy. We also want to investigate whether the progressive addition of adapters helps capture out-of-domain concepts, allowing the model to generalize better to unseen categories.

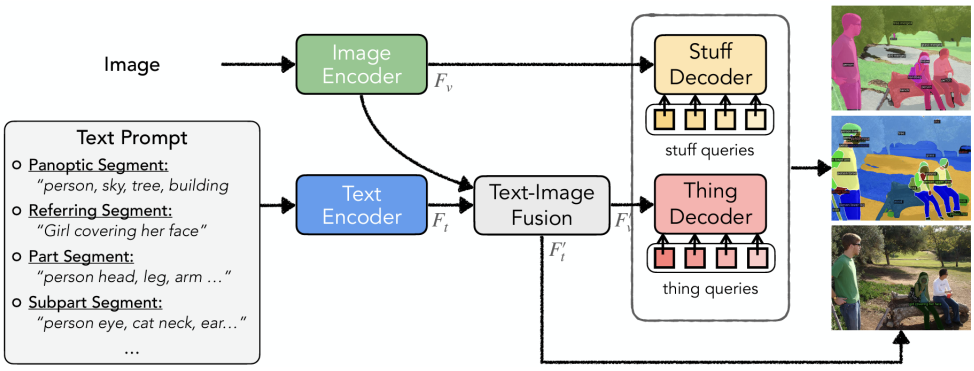


Figure 2.1: Diagram of HIPIE for hierarchical, universal and open-vocabulary image segmentation and detection. Adapted From [1].

2.2 Previous Literature Review

In preparation for this research, we have reviewed several foundational works that contribute to our understanding of open-vocabulary image segmentation:

2.2.1 CLIP (Contrastive Language-Image Pretraining)

CLIP is a foundational model that connects visual and textual representations using contrastive learning. By training on vast amounts of image-text pairs, CLIP enables zero-shot generaliza-

tion, allowing the model to classify images based on arbitrary textual descriptions without task-specific training. This concept is key for open-vocabulary segmentation, where textual descriptions are essential for segmenting unseen categories.

2.2.2 Segment Anything Model (SAM)

The Segment Anything Model (SAM) introduces the concept of promptability in segmentation tasks. SAM can generalize across tasks and datasets by using user-defined prompts for segmentation. Its ability to transfer zero-shot to new distributions aligns well with our goal of enabling open-vocabulary segmentation, as SAM demonstrates the power of flexible prompt-based interactions for identifying and segmenting unseen objects.

2.2.3 Vision Transformers (ViTs)

Vision Transformers (ViTs) represent a paradigm shift from traditional Convolutional Neural Networks (CNNs). ViTs treat images as sequences of patches and utilize self-attention mechanisms to process them. Their ability to capture long-range dependencies and hierarchical relationships makes them valuable in understanding multi-level segmentation tasks.

2.3 Next Steps

Moving forward, the main focus will be on extending HIPIE's capabilities by:

- Adding annotations for new categories, enhancing HIPIE's open-vocabulary understanding using NLP-based approaches.
- Incorporating adapters in both the Stuff and Thing Decoders to evaluate their impact on accuracy and generalization.
- Investigating whether additional adapters improve the model's ability to segment out-of-domain concepts.

This work will provide valuable insights into the effectiveness of hierarchical representations and multi-level segmentation in complex visual scenes.

2.4 Continued Literature Reviews

OMG-Seg: Is One Model Good Enough For All Segmentation? [2], *A Survey on Open-Vocabulary Detection and Segmentation: Past, Present, and Future* [3], *Osprey: Pixel Understanding with Visual Instruction Tuning* [4], *SED: A Simple Encoder-Decoder for Open-Vocabulary Semantic Segmentation* [5], *Scaling Open-Vocabulary Image Segmentation with Image-Level Labels* [6].

Bibliography

[1] X. Wang, S. Li, K. Kallidromitis, Y. Kato, K. Kozuka, and T. Darrell, "Hierarchical open-vocabulary universal image segmentation," 2023.

[2] X. Li, H. Yuan, W. Li, H. Ding, S. Wu, W. Zhang, Y. Li, K. Chen, and C. C. Loy, "Omg-seg: Is one model good enough for all segmentation?," 2024.

[3] C. Zhu and L. Chen, "A survey on open-vocabulary detection and segmentation: Past, present, and future," 2024.

[4] Y. Yuan, W. Li, J. Liu, D. Tang, X. Luo, C. Qin, L. Zhang, and J. Zhu, "Osprey: Pixel understanding with visual instruction tuning," 2024.

[5] B. Xie, J. Cao, J. Xie, F. S. Khan, and Y. Pang, "Sed: A simple encoder-decoder for open-vocabulary semantic segmentation," 2024.

[6] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," 2022.