Jialei Chen

Research Internship Application

jialeichen2021@gmail.com
+81-080-9730-5150
psmobile.github.io

名古屋大学
NAGOYA UNIVERSITY

Research Experience (Selected)

My research focuses on developing high-performance and efficient deep learning models for zero-shot and open-vocabulary semantic segmentation which aims to pixel-wisely understand an image and generalize to unlimited classes. The key outcomes of this work are summarized in the following publications:

Journal Publications

- **Frozen is Better than Learning: A New Design of Prototype-Based Classifier for Semantic Segmentation** (First Author)
  Pattern Recognition (IF 7.5/ JCR Q1)
  https://www.sciencedirect.com/science/article/pii/S0031320324001821

- **Uncertainty Teacher with Dense Focal Loss for Semi-Supervised Medical Image Segmentation** (First Author)
  Computers in Biology and Medicine (IF 7.0 / JCR Q1)
  https://www.sciencedirect.com/science/article/pii/S001048252200751X

- **Semantic Matters: A Constrained Approach for Zero-Shot Video Action Recognition** (Second Author, Provided Idea)
  Pattern Recognition (IF 7.5 / JCR Q1)
  https://www.sciencedirect.com/science/article/pii/S0031320325000627

- **Toward Explainable End-to-End Driving Models via Simplified Objectification Constraints** (Third Author, Supervisor as Second)
  IEEE Transactions on Intelligent Transportation Systems (IF 7.9 / JCR Q1)
  https://ieeexplore.ieee.org/abstract/document/10505932

Under Review / Preprints

- **Clip is Also a Good Teacher: A New Learning Framework for Inductive Zero-Shot Semantic Segmentation** (First Author)
  arXiv preprint submitted to *IEEE Transactions on Circuits and Systems for Video Technology*
  https://arxiv.org/pdf/2310.02296

- **Generalizable Semantic Vision Query Generation for Zero-Shot Panoptic and Semantic Segmentation** (First Author)
  arXiv preprint, submitted to *International Journal on Computer Vision*
  https://arxiv.org/pdf/2402.13697

- **Training-Free Open-Vocabulary Semantic Segmentation with Affinity Pyramid Refinement** (First Author)
  Submitted to *ICCV 2025*

- **BiXFormer: A Robust Framework for Maximizing Modality Effectiveness in Multi-Modal Semantic Segmentation** (First Author)
  Submitted to *ICCV 2025*

Experience and skill related to the applied topics

- **Programming**: Python, PyTorch, MMsegmentation
- **Development**: Model implementation, large-scale training, evaluation on benchmarks (ADE20K, COCO, etc.)
- **Vision-foundation models**: CLIP, DINO

Research Contents

My current research focuses on two core directions in semantic segmentation: **zero-shot and open-vocabulary segmentation**, and **multi-modal segmentation** to enable machines to recognize diverse objects under limited supervision and complex conditions. **Zero-shot and open-vocabulary segmentation** both aim to recognize the categories that do not exist during training. Zero-shot segmentation focuses on discovering more categories within the same dataset by learning from the relationships between seen and unseen classes. Open-vocabulary segmentation further enables recognition of new categories defined by text, allowing the model to adapt across different environments. Together, they help build segmentation models that are both scalable and adaptable without requiring extra annotations. For example, a suitcase may suggest someone is traveling, while a stroller may indicate childcare. Recognizing such objects helps the system interpret human activities. Even rare items like wheelchairs or walking aids can be crucial for understanding the behavior and needs of specific individuals, demonstrating the importance of my research contents.

Multi-modal segmentation in our research is not limited to image and language modalities, but also includes diverse sensor data such as LiDAR-based depth maps and point clouds from event cameras. This direction focuses on improving the robustness of segmentation under adverse conditions, such as low light or motion blur. By incorporating these additional modalities, the model can better perceive the environment and maintain accuracy where RGB input alone is insufficient. For example, on a rainy night, RGB images may fail to capture all objects due to low visibility, whereas LiDAR or event data can provide reliable signals to support accurate segmentation, showcasing the significance of utilizing multi-modality data.
Together, these two directions contribute to building flexible, scalable, and generalizable segmentation systems that support higher-level tasks such as human behavior understanding in real-world scenarios.

Experience and skill related to the applied topics

- Programming: Python, PyTorch, MMsegmentation
- Development: Model implementation, large-scale training, evaluation on benchmarks (ADE20K, COCO, etc.)
- Vision-foundation models: CLIP, DINO

Connection between my research and the research topic in cyberagent

The research topic I am interested in is “Multimodal Foundation Models for Human Behavior Understanding (人物行動理解のためのマルチモーダル基盤モデル),” with a focus on understanding how people behave by recognizing the objects around them. Accurately understanding human behavior is essential for building intelligent systems that can better serve society. For example, in a train station, a robot that recognizes someone carrying heavy luggage could proactively offer assistance; in autonomous driving, anticipating whether a pedestrian is about to cross the road is critical for safety. These scenarios all rely on the ability to perceive not just people, but also their interactions with objects in the environment. For example, if someone is holding a suitcase, they are probably traveling. If a person is pushing a shopping cart, they are likely shopping.

In real-world scenarios, although humans can easily recognize a wide variety of objects and understand how people interact with them, traditional segmentation models are typically restricted to a fixed set of categories defined during training. This limitation makes it difficult for such models to adapt to the diverse and ever-changing nature of everyday environments. To address this, my research focuses on **zero-shot semantic segmentation**, which enables machines to recognize categories that do not appear during training by leveraging semantic knowledge. This approach holds strong potential for real-world applications, especially in understanding complex human-object interactions.

Formally, zero-shot segmentation helps the model recognize more categories within a given dataset, even those not seen during training, by using shared meanings between classes. Open-vocabulary segmentation, on the other hand, helps the model adapt to new environments and recognize new classes using just text labels. By combining both, we can build models that are not only flexible with categories but also adaptable to different scenes, which is very important for understanding human behavior in the real world.

To support this kind of generalization, we also need to bring in other types of data, especially **text**. Language is how humans define and describe different objects. By adding language into the model, we can teach it to understand what objects are, even without many training examples. My work uses large vision-language models like CLIP to connect what the model sees with what we call things.

Based on this idea, I have developed segmentation frameworks that transfer CLIP’s knowledge to pixel-level prediction, allowing the model to detect many objects through text guidance. I also explore multi-modal designs that combine images with other sensor data, such as LiDAR or event cameras, so the model can handle difficult conditions like darkness or motion blur.

In short, my goal is to build a general system that can understand everything in a scene. Once we can recognize all objects accurately, we can then analyze how people behave by looking at what they are doing with those objects. By combining multi-modal inputs, zero-shot learning, and open-vocabulary ability, my research contributes to developing flexible and scalable foundation models for understanding human behavior, just as envisioned in 人物行動理解のためのマルチモーダル基盤モデル.

Academic Services

Reviewers for CVPR, IJCNN, IEEE Signal Processing Letters

Awards

- Frist class scholarship in Northeastern University 2019
- Second class scholarship in Northeastern University 2019

Language

Japanese: conversational, English: Fluent, Chinese: Native