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)

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 Author)
  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 centers around zero-shot and open-vocabulary semantic segmentation, with a particular focus on leveraging vision-language models such as CLIP for pixel-level understanding. This research is highly relevant to human behavior understanding, as accurate localization and recognition of both people and the objects they interact with are crucial.

- To address the challenge of recognizing unseen categories, I have developed methods that integrate CLIP-based semantic priors into segmentation models. For example, in our CLIP-to-Seg Distillation framework, we transfer both global and local vision-language knowledge from CLIP to segmentation backbones using distillation, without relying on CLIP during inference. This allows segmentation models to generalize to novel human-object interactions without additional annotations.

- Furthermore, we investigate semantic-centric alignment, where visual features are aligned with CLIP's semantic space rather than seen visual distributions, enabling robust recognition of novel actions and scenes. This approach is particularly valuable for behavioral understanding in open-world environments where seen and unseen categories coexist.

- In parallel, I am also exploring training-free approaches to probe the pixel-level vision-language matching probability for existing CLIP model. Our recent work integrates Affinity Pyramid Refinement into CLIP to capture hierarchical context and object coherence, enhancing segmentation accuracy without fine-tuning. This is beneficial for real-time deployment in dynamic environments such as surveillance and autonomous driving.

- In addition, I am actively studying multi-modal segmentation, particularly focusing on maximizing the contribution of modalities like LiDAR or event cameras under adverse conditions. This aligns with human behavior understanding in challenging environments, such as nighttime or occluded scenes.

- By combining zero-shot learning, semantic-centric alignment, training-free inference, and multi-modal integration, my research contributes to building flexible and robust segmentation systems that serve as foundational models for high-level human activity interpretation.

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🄵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.**

In everyday life, human behavior often depends on what objects people are interacting with. For example, if someone is holding a suitcase, they are probably traveling. If someone is carrying a shopping basket, they are likely shopping. These actions cannot be understood just by looking at their posture or movement—they are defined by the objects around them. So, to understand human behavior, we first need to understand what objects are in the scene and how people are interacting with them.

But in the real world, there are too many different objects to label them all by hand. It’s impossible to build a complete dataset for every category. To solve this, I study two key techniques: **zero-shot semantic segmentation** and **open-vocabulary segmentation**.

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—this 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 training-free methods and 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🄵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