Unsupervised Semantic Segmentation in Driving Scenario

清華 A122582

陳佑祥 You Xiang Chen

bob020416@gmail.com

***Abstract*— This project aims to adapt the CAUSAL Unsupervised Semantic Segmentation (CAUSE) framework to autonomous driving, addressing the need for accurate semantic segmentation without extensive labeled datasets. By integrating causal inference and self-supervised learning, the CAUSE framework categorizes semantic features and achieves precise pixel-level grouping in dynamic driving environments. We plan to enhance this methodology by employing advanced architectures and innovative training techniques across datasets like Cityscapes and Berkeley DeepDrive. Additionally, we will evaluate self-supervised pre-training models such as DINOv2 and MAE for their robust feature extraction. The goal is to develop a scalable, efficient unsupervised learning framework that enhances autonomous vehicle safety and functionality in urban settings.**

***Keywords*— Unsupervised Semantic Segmentation, Driving Scenario**

I. Introduction

Semantic segmentation is pivotal in advancing autonomous driving technologies, having benefited significantly from deep neural networks (DNNs) and large-scale datasets over the past decade. However, the demand for pixel-level annotations presents notable challenges, especially in the dynamic and complex scenarios typical of driving environments.

Conventional supervised methods are often prohibitively expensive and time-consuming, prompting a shift towards unsupervised semantic segmentation (USS) techniques. These methods exploit visual consistency and multi-view equivalence but have not been fully tailored to the intricacies of driving scenes.

Our project seeks to adapt and expand the CAUSAL Unsupervised Semantic Segmentation (CAUSE) framework, originally designed for general applications, to the unique conditions of driving scenes. The CAUSE methodology utilizes a two-step process involving the construction of a concept clusterbook to categorize potential semantic categories and leveraging this framework for pixel-level grouping via self-supervised learning. This integration of causal inference addresses the critical challenge of clustering without supervision.

We plan to implement the CAUSE framework across diverse driving scene datasets like Cityscapes and Berkeley DeepDrive and explore various self-supervised pre-training models, including DINOv2 and MAE. These models are chosen for their ability to extract robust, discriminative features essential for accurately representing the complexities of urban environments.

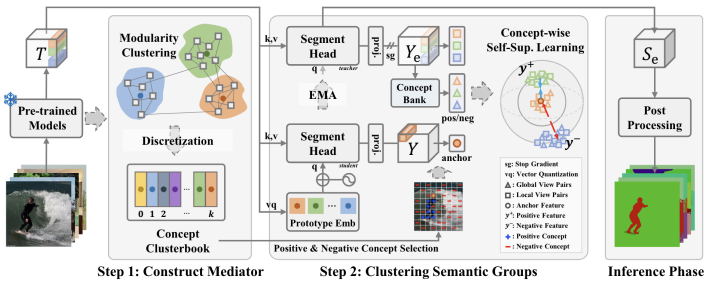

Figure 1. CAUSE Method

Our goal is to develop a more flexible and effective unsupervised learning framework that can reduce reliance on extensive labeled datasets while still providing the high granularity and accuracy needed for autonomous driving applications.

II. Related Work

Unsupervised Semantic Segmentation (USS) has evolved as an alternative to labor-intensive annotated segmentation methods. Early efforts, such as Invariant Information Clustering (IIC) by Ji et al. (2019), laid the groundwork by maximizing mutual information from augmented views. This approach has been expanded by incorporating inductive biases for cross-image correspondences and saliency information, notably by Hwang et al. (2019) and Van Gansbeke et al. (2021).

Recent advancements have leveraged

self-supervised learning frameworks to utilize pre-trained features for USS, with notable contributions from Caron et al. (2021) and Hamilton et al. (2022). These methods have improved segmentation quality by using self-supervised representations as pseudo labels and integrating additional prior knowledge. However, challenges remain due to the lack of well-defined clustering targets in USS, leading to potential suboptimal segmentation outcomes.

III. METHODOLOGY

The CAUSE framework utilizes a two-step process, integrating causal inference to address the challenges of unsupervised semantic segmentation.

3.1 Constructing Concept Clusterbook

The initial phase focuses on creating a mediator, termed as the concept clusterbook M. This mediator:

- Captures and categorizes potential semantic categories from pre-trained model features T using modularity maximization.

- Enhances the clarity and usability of features by transforming them into a discretized form suitable for segmentation.

Algorithm 1 (STEP 1) Maximizing Modularity for Constructing Concept Clusterbook M

Require: Image Samples $X \sim$ Data, Pre-trained Model f, Concept Fractions $M \in \mathbb{R}^{k \times c}$
Initialize M
for $X \sim$ Data **do**
  $T \in \mathbb{R}^{hw \times c} \leftarrow f(X)$ ▷ Pre-trained Model Representation
  $\mathcal{A} \leftarrow \max(0, \cos(T, T)) \in \mathbb{R}^{hw \times hw}$ ▷ Affinity matrix
  $d, e \leftarrow \mathcal{A}$ ▷ Degree Matrix and Number of Total edges
  $\mathcal{C} \leftarrow \max(0, \cos(T, M)) \in \mathbb{R}^{hw \times k}$ ▷ Cluster Assignment Matrix
  $\mathcal{H} \leftarrow \frac{1}{2e}\mathrm{Tr}\left(\tanh\left(\frac{\mathcal{C}\mathcal{C}^T}{\tau}\right)\left[\mathcal{A} - \frac{dd^T}{2e}\right]\right)$ ▷ Maximizing Modularity ($\tau = 0.1$)
  $M \leftarrow \text{Increase}(\mathcal{H})$ ▷ Updating Concept ClusterBook (lr: 0.001)
end for

FIGURE 2. ALGORITHM1

3.2 Enhancing Likelihood of Semantic Groups through Self-Supervised Learning

Once the mediator is established, the next step involves:

- Employing self-supervised learning techniques to refine the segmentation outputs.

- Using concept-wise comparisons to strengthen the model's ability to differentiate between semantic categories effectively.
- Applying a segmentation head to map the discretized features back to pixel-level semantic labels accurately.

Algorithm 2 (STEP 2): Enhancing Likelihood of Semantic Groups through Self-Supervised Learning

Require: Head S;θ_S, Head-EMA S_{ema};$\theta_{S_{\text{ema}}}$, Clusterbook M, Distance $\mathcal{D}_M$, Concept Bank Y_{bank}
for $X \sim$ Data **do**
  $T \leftarrow f(X)$ ▷ Pre-trained Model Representation
  $Q \leftarrow T$ ▷ Vector Quantization from M
  $Y, Y_{\text{ema}} \leftarrow S(Q, T), S_{\text{ema}}(Q, T)$ (∗ MLP: $S(T), S_{\text{ema}}(T)$) ▷ Segmentation Head Output
  $y \sim Y$ ▷ Anchor Selection (Appendix B for Detail)
  $y^+, y^- \sim \{Y_{\text{ema}}, Y_{\text{bank}} \mid y\}$ ▷ Positive/Negative Selection from $\mathcal{D}_M$ (Appendix B for Detail)
  $p \leftarrow \mathbb{E}_y\left[\log \mathbb{E}_{y^+}\left[\frac{\exp(\cos(y,y^+)/\tau)}{\exp(\cos(y,y^+)/\tau)+\sum_{y^-}\exp(\cos(y,y^-)/\tau)}\right]\right]$ ▷ Self-supervised Learning
  $\theta_S \leftarrow \text{Increase}(p)$ ▷ Updating Parameters of Segmentation Head (lr: 0.001)
  $\theta_{S_{\text{ema}}} \leftarrow \lambda\theta_{S_{\text{ema}}} + (1-\lambda)\theta_S$ ▷ Exponential Moving Average ($\lambda : 0.99$)
  $Y_{\text{bank}} \leftarrow \mathbf{R}^2(Y_{\text{bank}}, Y_{\text{ema}})$ ▷ $\mathbf{R}^2$: **R**andom Cut Y_{bank} and **R**andom Sample Y_{ema}
end for

FIGURE 3. ALGORITHM2

3.3 Proposed Improvements and Experiments

To extend the capabilities, we propose several improvements:
1. Integration of Advanced Architectures: Experimenting with hybrid models combining CNNs for feature extraction and Transformers for global context integration, might enhance the segmentation accuracy in dynamic driving scenes.

2. Innovative Training Techniques: Testing different self-supervised learning paradigms, such as contrastive learning with more sophisticated negative sampling strategies and triplet loss, to improve the robustness and discrimination capacity of the model.

3. Experimentation Across Datasets: Applying the modified CAUSE framework on diverse datasets tailored to driving scenarios, such as Cityscapes and Berkeley DeepDrive, to assess and refine its performance across various urban settings.

4. Utilization of Enhanced Pre-training Models: Exploring the impact of newer self-supervised pre-training models like DINOv2 and MAE, which could provide more nuanced feature representations beneficial for complex segmentation tasks encountered in autonomous driving.

IV. CONCLUSION

This project aims to significantly advance the field of autonomous driving by adapting the CAUSAL Unsupervised Semantic Segmentation (CAUSE) framework to meet the unique challenges of driving scenes. By leveraging advanced self-supervised learning techniques and integrating causal inference, we anticipate developing a robust, unsupervised semantic segmentation system capable of operating effectively in dynamic urban environments without reliance on extensively labeled datasets.

References

[1] Wang, X., Ma, H., & You, S. (2020). Deep clustering for weakly-supervised semantic segmentation in autonomous driving scenes. Neurocomputing, 381, 20-28.

[2] Vobecky, A., Hurych, D., Simeoni, O., Gidaris, S., Bursuc, A., Perez, P., & Sivic, J. (2022). Drive&Segment: Unsupervised Semantic Segmentation of Urban Scenes via Cross-modal Distillation. valeo.ai, Czech Institute of Informatics, Robotics and Cybernetics, CTU in Prague.

[3] Kalluri, T., Varma, G., Chandraker, M., & Jawahar, C. V. (2019). Universal Semi-Supervised Semantic Segmentation. Center for Visual Information Technology, IIIT Hyderabad and University of California, San Diego.