Literature Review: Self-Supervised Learning and World Models for Autonomous Driving

Tejdeep Chippa
Tandon School of Engineering
New York University
tc4263@nyu.edu

Venkat Kumar Laxmi Kanth Nemala
Tandon School of Engineering
New York University
vn2263@nyu.edu

Abstract

This project adapts VSLNet for the Ego4D Goal-Step challenge, enhancing action localization in untrimmed, first-person videos using text descriptions. Leveraging pre-trained LaViLa for contextual embeddings, we integrate stacked visual encoders (optimal with two encoders at 11.86 mIoU@0.3) and vision-enhancer attention to refine context-query alignment. These enhancements improve localization accuracy while balancing precision and memory efficiency.

1 Introduction

In recent years, self-supervised learning (SSL) has revolutionized machine learning by enabling models to learn useful representations from raw data without large labeled datasets. Unlike traditional supervised learning, which relies on extensive human annotation, SSL discovers patterns and structures within the data itself, making it highly scalable and adaptable across domains. This approach has been particularly transformative in computer vision and autonomous driving, where acquiring labeled data is both time-consuming and expensive.

A critical advancement in SSL is the development of world models, which aim to create structured representations of an environment. In the context of autonomous driving, a world model enables a self-driving system to predict, interpret, and navigate its surroundings more effectively by learning spatial and temporal relationships. However, most existing SSL techniques fall into two main categories: contrastive learning, which relies on handcrafted positive and negative pairs, and generative learning, which focuses on reconstructing missing data. Both approaches have limitations in capturing high-level semantic features efficiently.

To address these challenges, the Joint-Embedding Predictive Architecture (JEPA) introduces a novel approach by predicting abstract representations rather than raw pixel reconstructions. JEPA-based models, such as I-JEPA for images and AD-L-JEPA for LiDAR data, improve efficiency and generalization in learning semantic representations. Unlike contrastive learning, JEPA eliminates the need for explicitly defined data augmentations, making it more adaptable to structured environments. This literature review explores their impact on autonomous driving applications, particularly in self-supervised pretraining for steering control in CARLA.

2 Image-based JEPA (I-JEPA)

Building on the Joint-Embedding Predictive Architecture (JEPA), I-JEPA extends its principles to image-based self-supervised learning, focusing on predicting abstract representations rather than reconstructing raw pixels. This approach significantly enhances semantic feature learning, making it more efficient than traditional generative or contrastive methods. Unlike contrastive learning, which relies on handcrafted augmentations and positive-negative pairs, I-JEPA uses a multi-block masking

strategy to predict missing representations from a single context block, improving scalability and adaptability.

The model achieves a top-1 accuracy of 79.3

Additionally, the exponential moving average (EMA) update of the target encoder stabilizes training and prevents representation collapse. By learning high-level semantic features in embedding space, rather than reconstructing raw image pixels, I-JEPA improves transfer learning performance, making it an ideal solution for scalable vision-based applications such as autonomous driving.

3 AD-L-JEPA: Self-Supervised Spatial World Models for LiDAR-Based Autonomous Driving

Building on the Joint-Embedding Predictive Architecture (JEPA), AD-L-JEPA extends its principles to LiDAR-based autonomous driving, introducing a self-supervised spatial world model that enhances 3D perception. Unlike traditional approaches that reconstruct raw point clouds, AD-L-JEPA predicts representations in Bird's Eye View (BEV) space, allowing for a more structured and efficient spatial understanding of driving environments.

A core innovation of AD-L-JEPA is its Modified BEV-Guided Masking strategy. Unlike existing masking methods that only mask occupied areas, this approach masks both empty and non-empty regions, ensuring the model learns a more complete spatial representation. This helps the model handle occlusions, where objects like cars may be hidden from direct LiDAR visibility but can still be inferred from contextual information.

By operating in the representation space rather than reconstructing raw data, AD-L-JEPA extracts high-level geometric and semantic features, significantly improving 3D object detection and scene understanding. The method also demonstrates improved label efficiency, outperforming existing self-supervised approaches while requiring less annotated data for fine-tuning. Additionally, unlike some SSL techniques that degrade performance during transfer, AD-L-JEPA consistently shows positive transfer across datasets, enhancing generalization and robustness in real-world scenarios.

Empirical evaluations demonstrate AD-L-JEPA's efficiency, achieving 5× faster pre-training than generative methods like Occupancy-MAE and outperforming contrastive and generative approaches on datasets like Waymo and KITTI3D. These advantages position AD-L-JEPA as a scalable and adaptable solution for self-supervised learning in autonomous driving applications.

4 Results Discussions

The performance of I-JEPA and AD-L-JEPA demonstrates the strength of JEPA-based self-supervised learning in both image-based and LiDAR-based tasks, significantly outperforming traditional contrastive and generative self-supervised learning approaches. These models provide improvements in computational efficiency, label efficiency, and transfer learning, making them particularly well-suited for autonomous driving applications.

I-JEPA achieves a top-1 accuracy of 79.3% on ImageNet-1K with a ViT-H/14 model, outperforming Masked Autoencoders (MAE) (77.2%) while requiring 5× fewer training iterations than contrastive methods. Unlike DINO and iBOT, which rely on handcrafted augmentations, I-JEPA eliminates such dependencies, improving adaptability across datasets. Its efficiency is particularly evident in semi-supervised learning, where it achieves 69.4% top-1 accuracy on 1% labeled ImageNet, outperforming MAE by 1.3 percentage points (68.0%). These results confirm JEPA's ability to extract high-level semantic representations with minimal supervision, making it an efficient and scalable alternative to traditional SSL methods.

AD-L-JEPA achieves state-of-the-art 3D object detection, with a mean Average Precision (mAP) of 67.92% on KITTI3D, outperforming Occupancy-MAE (66.63%) and models without self-supervised pretraining (66.36%). It also requires 50% fewer labeled samples to reach the same performance as supervised models, demonstrating superior label efficiency. Additionally, AD-L-JEPA achieves 5× faster pre-training compared to Occupancy-MAE, significantly reducing computational costs. A key factor in its success is Modified BEV-Guided Masking, which enhances occlusion reasoning by learning spatial relationships in Bird's Eye View (BEV) space, making it highly effective for

real-world self-driving applications. Unlike contrastive and generative LiDAR SSL approaches, it learns object-centric embeddings rather than relying on raw reconstruction, improving transferability and robustness across datasets.

Model	Domain	Pretraining Speed	Top-1 Accuracy / mAP	Key Advantage
I-JEPA	Image (ImageNet-1K)	5× faster than contrastive SSL	79.3% (ViT-H/14)	No handcrafted augmentations
I-JEPA (1% labels)	Image (ImageNet-1K)	Efficient for semi-supervised learning	69.4% (ViT-L/16)	Outperforms MAE (+1.4%)
AD-L-JEPA	LiDAR (KITTI3D)	5× faster than Occupancy-MAE	67.92% (mAP)	BEV-guided masking for occlusion handling
AD-L-JEPA (Transfer)	LiDAR (Waymo $\rightarrow$ KITTI3D)	Retains performance across datasets	+1.3% over baseline	No negative transfer

Table 1: Comparison of I-JEPA and AD-L-JEPA Models

5 Implications for Autonomous Driving

The combination of I-JEPA and AD-L-JEPA presents a powerful approach for self-supervised learning in autonomous driving, particularly for steering control in CARLA. I-JEPA's ability to learn high-level semantic features from images makes it well-suited for road structure recognition, lane detection, and object identification, all crucial for visual-based navigation. Meanwhile, AD-L-JEPA's BEV-based spatial modeling enhances depth perception, occlusion reasoning, and object localization, improving the vehicle's ability to handle complex, dynamic environments.

By integrating I-JEPA's visual representation learning with AD-L-JEPA's LiDAR-based spatial reasoning, a multi-modal self-driving system can be trained with minimal labeled data and fine-tuned efficiently for adaptive steering control. This approach would allow the model to generalize across different driving conditions, reducing domain shift issues and improving real-time decision-making in CARLA.

6 Conclusion and Brief Project Approach

The integration of self-supervised learning (SSL) with world models is transforming autonomous driving by reducing reliance on labeled data, improving real-time perception, and enhancing model generalization. The results of I-JEPA and AD-L-JEPA demonstrate how JEPA-based architectures effectively learn high-level semantic and spatial representations without requiring explicit augmentations or pixel reconstructions. I-JEPA's image-based learning enhances scene understanding, while AD-L-JEPA's LiDAR-based BEV modeling improves depth perception and occlusion handling, making them highly valuable for multi-modal self-driving applications.

Building on these insights, a potential project could focus on integrating I-JEPA and AD-L-JEPA to create a multi-modal self-driving system that fuses image and LiDAR inputs for improved scene understanding and steering control in CARLA. Optimizing computational efficiency would enable real-time deployment in autonomous vehicles, while reinforcement learning (RL) fine-tuning could further enhance adaptive steering and decision-making. By leveraging self-supervised world models, this approach aims to make autonomous systems more scalable, robust, and capable of handling diverse driving environments.