Multi-modal Multi-task Foundation Models for Road Scene Understanding

Image
Video
Text
LiDAR
Action
Audio

Multi-modal Input

Multi-task Output

Perception
Prediction
Planning
Reasoning
VQA
Captioning

Prerequisites

Basic Architecture

Multi-modal Data

Task 1
Task 2
Task 3
Multi-task Learning

Multi-modal Learning

Target
Target
Pretraining Objectives

Layer N
Layer 2
Layer 1
new data
new task
Fine-tuing Techniques

Common Practices

Task-specific Models	Unified Multi-task Models	Unified Multi-modal Models	Prompting Foundation Models
—From instance-level to global-level —From closed-set to open-set —From single-modal to multi-modal	—Task-specific outputs —Unified language outputs	—LLM functions as sequence modeling —Cross-modal interaction in VLM	—Textual prompt —Visual prompt —Multi-step prompt —Task-specific prompt —Geometric prompt —Prompt pool

Advanced Models

Known
Unknown
Towards Open-world Understanding

Source Domain
Adaptation
Distillation
Target Domain
Efficient Transfer for Road Scenes

Search
Store
Knowledge base
Continual Learning

Recoery
Interact
Learn to Interact

Prompt: car
AI Generate
Generative Foundation Model

Challenges & Trends

CARLA
Closed-Loop System

Interpretability

Low-resource Condition

Emodied Driving Agent

World Model