Towards Unified Perception: End-to-End Scene Understanding Models for Autonomous Driving

Sowmiya Narayanan Govindaraj
Senior Machine Learning Engineer

AGENDA

What We'll Cover

1. The Fragmentation Problem
2. The Unified Perception Paradigm
3. Multi-Sensor Fusion Transformers
4. Real-World Deployment Challenges
5. Foundation Models for Autonomous Driving

INTRODUCTION

The Fragmentation Problem in Traditional Perception

Conventional perception architectures treat object detection, semantic segmentation, and depth estimation as isolated tasks. This modular approach introduces redundant computation across separate neural networks and propagates errors between pipeline stages.

Each module processes identical sensor streams independently, multiplying inference latency and complicating deployment. The lack of shared representations prevents the system from learning cohesive scene understanding.

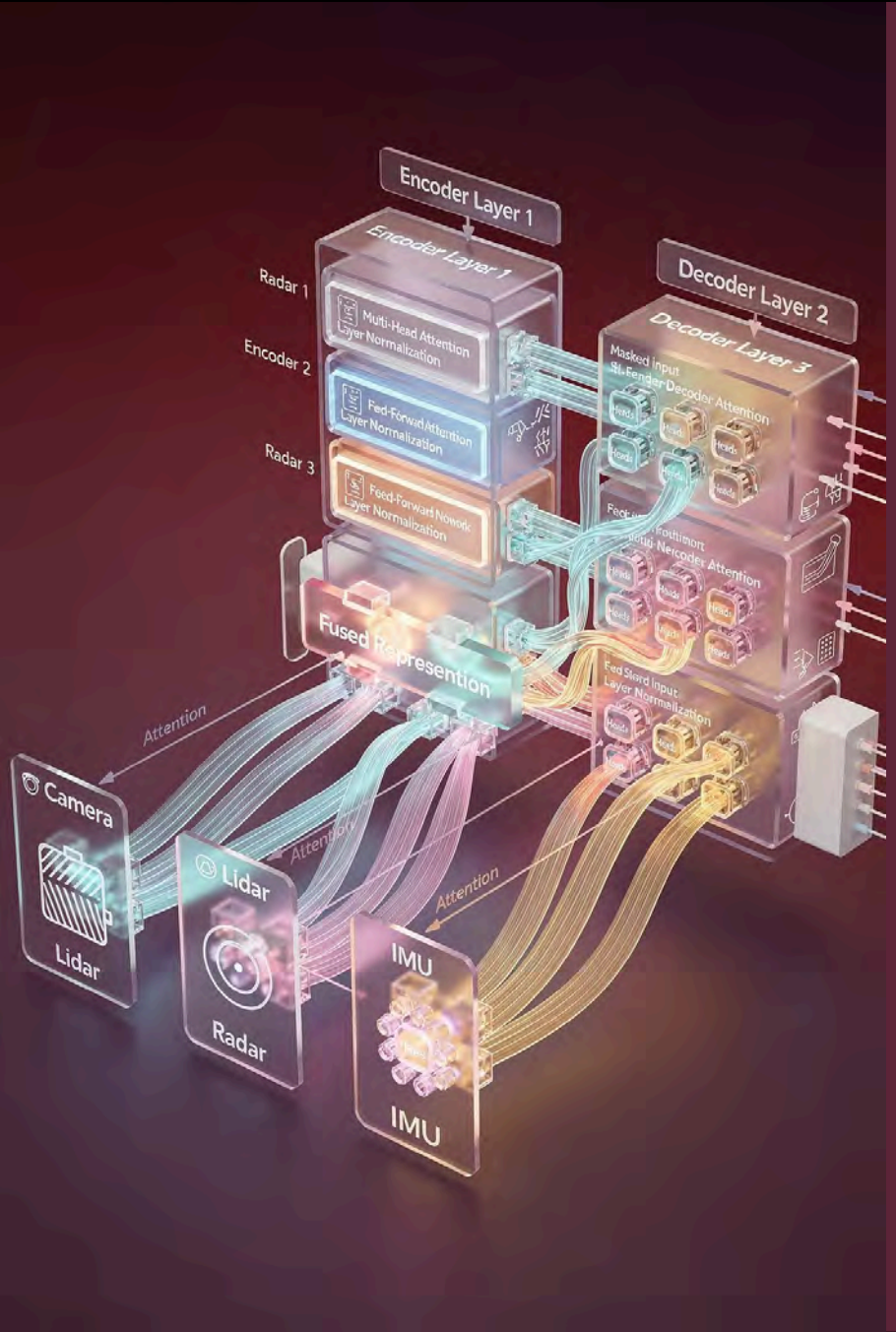
The Unified Perception Paradigm

Single Integrated System

Joint learning of spatial, semantic, and temporal representations within one deep learning architecture

Unified perception consolidates multiple perception tasks into a single end-to-end model. By learning shared feature representations, these systems eliminate redundancy and enable cross-task knowledge transfer.

The architecture processes multi-modal sensor inputs LiDAR point clouds, camera imagery, radar returns through shared encoders before branching into task-specific heads. This approach reduces computational overhead whilst improving accuracy through multi-task optimisation.



ARCHITECTURE

Multi-Sensor Fusion Transformers

Vision transformers adapted for sensor fusion leverage self-attention to model long-range spatial dependencies. Cross-modal attention layers align features from heterogeneous sensors camera RGB data with LiDAR geometry enabling the network to learn complementary representations. Token-based architectures naturally handle variable-resolution inputs and facilitate end-to-end optimisation across detection, segmentation, and depth tasks simultaneously.

Core Architectural Components

Multi-Task Feature Decoupling

Shared backbone extracts common representations whilst task-specific branches learn specialised features for detection, segmentation, and depth

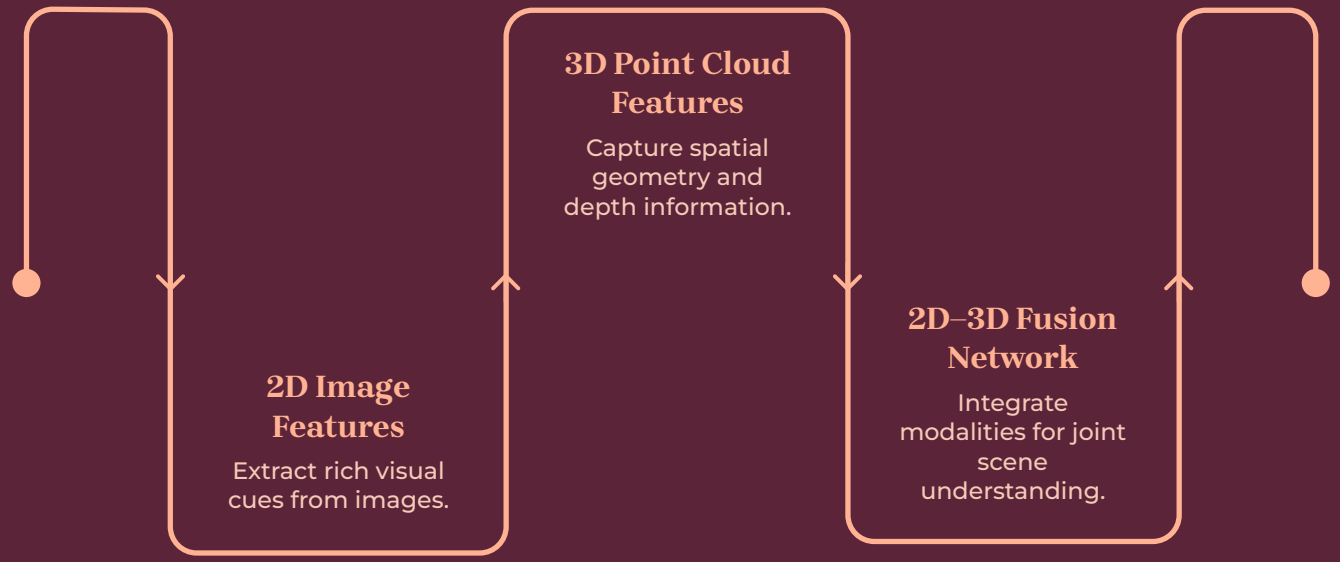
Cross-Modal Alignment Networks

Projection layers and contrastive learning align LiDAR geometric features with camera appearance information in a unified latent space

Temporal Fusion Modules

Recurrent or temporal attention mechanisms aggregate sequential observations to model motion and maintain scene continuity

Joint 2D–3D Learning Framework



Integration Benefits

- Accurate 3D bounding box prediction from camera inputs
- Dense semantic labels projected into point clouds
- Depth-aware object detection with uncertainty quantification
- Unified feature space for downstream planning modules

PRODUCTION INSIGHTS

Adaptive Attention Mechanisms

Learnable attention gates dynamically weight contributions from different sensors based on scene context and sensor reliability. In adverse weather, the network automatically upweights LiDAR whilst downweighting degraded camera inputs. This adaptive fusion proves essential for robust perception across diverse operational design domains encountered in production deployments at scale.

Scaling Fusion Pipelines in Real-World Environments

Weather Variation

Rain, fog, and snow degrade sensor performance non-uniformly. Multi-modal fusion with weather-aware attention provides graceful degradation rather than catastrophic failure.

Domain Shifts

Geographic variation in road infrastructure, signage, and driving behaviour causes distribution shift. Domain adaptation techniques and continual learning protocols maintain performance across regions.

Long-Tail Scenarios

Rare events—construction zones, emergency vehicles, unusual pedestrian behaviour—require targeted data collection and balanced sampling strategies during training to prevent model underperformance.

Deployment Complexity Reduction

Traditional Modular Stack

- Separate models for detection, segmentation, depth
- Multiple inference passes through distinct networks
- Complex inter-module interfaces and error propagation
- Higher latency and computational overhead

Unified Perception System

- Single end-to-end model handling all tasks
- One forward pass for complete scene understanding
- Direct optimisation of final perception metrics
- Reduced latency and simplified deployment pipeline

EMERGING FRONTIERS

Self-Supervised Fusion Methods

Contrastive multi-modal learning enables models to learn alignment between camera and LiDAR without explicit supervision. Temporal consistency constraints and cross-sensor reconstruction objectives provide training signals from unlabelled data. These approaches reduce annotation costs whilst scaling to massive datasets, critical for capturing long-tail distributions in autonomous driving scenarios.

Domain-Adaptive Perception

Transferring perception models across operational domains requires specialised adaptation techniques. Adversarial domain alignment minimises distribution mismatch between source and target domains in latent feature space.

Test-time adaptation protocols adjust batch normalisation statistics and model parameters online as the vehicle encounters new environments. Meta-learning frameworks enable rapid fine-tuning with limited target-domain data, accelerating geographical expansion.

Foundation Models for Autonomous Driving

- **Pre-Trained Representations**

  Models trained on millions of driving hours learn general-purpose visual priors, reducing downstream task-specific data requirements

- **Multi-Task Backbone**

  Shared encoder supports detection, tracking, prediction, and planning through task-specific decoder heads with minimal fine-tuning

- **Zero-Shot Generalisation**

  Foundation models exhibit surprising capability on unseen object categories and scenarios without explicit training examples

KEY ADVANTAGES

Impact on Modern Autonomy Systems

- **Latency Reduction**

  Single forward pass eliminates redundant computation across modular perception tasks

- **Accuracy Improvement**

  Shared representations and multi-task learning enhance performance through joint optimisation

- **Deployment Simplification**

  Unified architecture reduces system complexity and maintenance overhead in production fleets

Accelerating Safer Autonomous Systems

Unified perception architectures directly contribute to safety by providing consistent, coherent scene understanding. Eliminating error propagation between modules reduces failure modes. Joint training enables the model to resolve ambiguities by leveraging complementary sensor modalities.

Temporal reasoning improves prediction of dynamic agents whilst adaptive fusion maintains robustness under sensor degradation. These capabilities translate to measurable safety improvements in disengagement rates and critical intervention frequencies across production autonomous driving systems.

CONCLUSION

The Future of Perception for Autonomy

Unified architectures are transforming perception

End-to-end models streamline pipelines and improve system robustness

Multi-sensor fusion is essential

Adaptive attention and cross-modal alignment enable resilient scene understanding

Foundation models will accelerate progress

Pre-trained representations and zero-shot capabilities reduce development cycles

As unified perception continues to mature, autonomous systems will become safer, more scalable, and more capable of handling the complexity of real-world driving environments.

Thank You!

Sowmiya Narayanan Govindaraj || Conf42 Machine Learning 2026