

Capstone Project Report

Uncertainty-Aware Cooperative Occupancy Prediction with Attention Mechanisms

by

Harshini Murugadoss

Submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in

Robotics

UNIVERSITY OF CALIFORNIA, RIVERSIDE

December 22, 2025

Committee:

Jiachen Li, Chair

Hang Qiu

Abstract

As autonomous mobile robots transition from controlled industrial settings to dynamic, shared human environments such as hospitals, airports, and warehouses ensuring safe and efficient navigation becomes a fundamental challenge. A key limitation of single-robot perception systems is their restricted line-of-sight: static obstacles like walls or shelves create occlusions that hide dynamic hazards such as pedestrians. These “blind spots” prevent robots from anticipating collisions, limiting both safety and operational speed. This project addresses these challenges by developing an Enhanced Early Fusion framework for cooperative perception, enabling teams of robots to jointly reason about the future state of their environment.

We focus on the problem of Uncertainty-Aware Occupancy Grid Prediction. Our approach fuses raw LiDAR data from multiple localized agents to generate a unified spatiotemporal forecast of the scene. The core contribution is the integration of a Dual Attention Mechanism into a Recurrent Variational Autoencoder (RVAEP) backbone. While standard fusion methods treat all incoming data equally, often producing noisy predictions, the proposed attention modules selectively emphasize informative features. A Spatial Attention Module focuses computation on dynamic regions such as moving pedestrians, while a Channel Attention Module adaptively weights feature maps to suppress sensor noise and alignment artifacts.

The system is trained and evaluated in a high-fidelity Unity3D simulation featuring complex indoor layouts and unpredictable human motion. Using a dataset of over 360,000 sequences, experiments show that the attention-enhanced model significantly outperforms standard fusion baselines. In particular, it achieves an 11.8% reduction in Binary Cross Entropy (BCE) loss, along with improvements in Structural Similarity (SSIM) and Intersection-over-Union (IoU). By producing both accurate occupancy forecasts and calibrated entropy-based uncertainty maps, the proposed framework provides a strong foundation for safety-aware multi-robot navigation in human-centered environments.

1 Introduction

1.1 Motivation

The field of autonomous mobile robotics is currently undergoing a paradigm shift. Historically, robots were confined to highly controlled, deterministic environments such as factory floors or automated warehouses, where human presence was minimal and strictly regulated. Today, however, there is a growing demand for robots to operate in “social” environments dynamic, cluttered spaces shared with humans, such as busy hospitals, school campuses, airport terminals, and pedestrian plazas Li et al. [2024]. In these settings, the primary operational directive shifts from pure efficiency to absolute safety and social compliance.

A fundamental challenge in social navigation is the “Perceptual Horizon” problem. A single robot, no matter how sophisticated its sensors, is physically constrained by its line-of-sight. On-board sensors like LiDAR or depth cameras cannot penetrate solid objects. In a crowded corridor, a wall, a bookshelf, or even another slow-moving vehicle can create significant “blind spots.” These occlusions are not merely inconveniences; they are safety hazards. A robot turning a blind corner or overtaking a stopped obstacle may find itself on a collision course with a pedestrian emerging from the occlusion. This limitation forces single autonomous agents to move conservatively often stopping or crawling at inefficient speeds because they cannot confidently predict the state of the environment beyond what they immediately see. To achieve human-level fluidity and safety, robots need a way to see the “unseen.”

1.2 The Cooperative Approach

The solution to the single-agent perceptual limit lies in Cooperative Perception, enabled by Vehicle-to-Vehicle (V2V) communication Bai et al. [2024]. This paradigm treats a fleet of robots not as isolated individuals, but as a distributed sensor network. If Robot A is blocked by a wall but Robot B has a clear line-of-sight to the area behind that wall, Robot B can share its sensory information with Robot A. Through this collaboration, Robot A effectively gains “X-ray vision,” allowing it to anticipate and plan for dynamic actors that are physically invisible to its own sensors Wang et al. [2025].

However, implementing this cooperation is non-trivial. It is not enough to simply broadcast raw data to every nearby agent. Naive data aggregation faces two major hurdles:

1. **Bandwidth and Latency:** Transmitting high-resolution sensor streams (such as point clouds or images) consumes significant bandwidth, which is a scarce resource in wireless robot networks.

2. **Noise and Misalignment:** Data from different robots is often collected from different viewpoints and may suffer from localization errors (e.g., GPS drift or odometry slip). Fusing slightly misaligned maps can result in “ghosting” artifacts, where a single object appears duplicated or static walls seem to drift.

Therefore, a robust cooperative system requires *intelligent fusion*, a mechanism that does not simply stack data together, but actively reasons about which information is reliable and which constitutes noise.

1.3 Project Objective & Contributions

The primary objective of this capstone project is to develop an end-to-end framework for Uncertainty-Aware Cooperative Occupancy Prediction. We aim to build a deep learning system that allows a team of robots to seamlessly fuse their LiDAR observations to generate a high-fidelity, predictive map of their surroundings. Crucially, this system is designed to be introspective not only predicting *where* obstacles will be in the future, but also quantifying *how confident* it is in those predictions.

This work makes the following technical contributions to the field of multi-robot perception:

1. **Unity3D-Based Data Pipeline:** We developed a high-fidelity simulation environment in Unity3D that bridges the gap between abstract algorithms and realistic physics. Unlike simple grid-world simulators, our platform models realistic LiDAR ray-casting, vehicle dynamics, and NavMesh-driven human crowd behavior, providing a rich and challenging dataset for training.
2. **Attention-Enhanced Neural Architecture:** We propose a novel architecture that integrates Dual Attention Mechanisms into a standard Recurrent Variational Autoencoder (RVAEP).
 - **Spatial Attention:** Acts as a “visual spotlight,” enabling the network to identify and emphasize important regions (e.g., dynamic actors) while suppressing static background noise.
 - **Channel Attention:** Acts as a “feature filter,” dynamically weighting internal feature maps to prioritize motion- and occupancy-related signals, effectively reducing sensor artifacts common in multi-robot fusion.
3. **Uncertainty-Aware Evaluation:** Beyond standard accuracy metrics, we evaluate the model’s ability to produce calibrated uncertainty estimates. We demonstrate that the attention-based architecture not only improves structural accuracy (measured by SSIM) but also reduces prediction entropy in well-observed regions, providing downstream navigation planners with a reliable distinction between “safe” and “unsafe” zones.

2 Methodology

Our proposed framework tackles the multi-agent occupancy prediction problem using an Attention-Enhanced Early Fusion paradigm. This section details the end-to-end pipeline, from data ingestion to the final probabilistic forecast.

2.1 Early Fusion Strategy

In cooperative perception, “fusion” is the process of combining sensor data from multiple spatially distributed agents. We adopt an Early Fusion strategy, which integrates information at the raw data level rather than the feature or decision level. This approach preserves the maximum amount of spatial information, which is critical for detecting small or partially occluded objects.

For a team of N robots, each robot i receives the raw LiDAR point clouds P_j from its neighbors $j \in N$.

1. **Coordinate Transformation:** Before fusion, neighbor point clouds must be aligned to the ego-robot’s frame. Using the known relative pose (transformation matrix $T_{j \rightarrow i}$), we project all points into a common coordinate system:

$$P'_j = T_{j \rightarrow i} \cdot P_j$$

2. **Voxelization:** The aggregated point cloud is then discretized into a 2D Occupancy Grid Map (OGM) of size $H \times W$ (specifically 64×64). Each grid cell is assigned a binary state: *Occupied* (1) if it contains LiDAR points, or *Free* (0) otherwise.
3. **Temporal Stacking:** To capture motion dynamics, we do not process a single frame in isolation. Instead, we stack the past T frames (where $T = 5$) along the channel dimension, creating a spatiotemporal tensor

$$X_t \in \mathbb{R}^{T \times H \times W}.$$

This tensor serves as the rich input context for our deep learning model.

2.2 Baseline Architecture: SOGMP++ (RVAEP)

Our backbone architecture is based on the Self-Supervised Occupancy Grid Map Prediction (SOGMP) model Xie and Dames [2023], which utilizes a Recurrent Variational Autoencoder (RVAEP) structure. This architecture is designed to handle two fundamental properties of the environment: dynamic motion and future uncertainty.

- **ConvLSTM Encoder:** Standard Convolutional Neural Networks (CNNs) are effective at spatial feature extraction but lack temporal memory. We employ a ConvLSTM network Toyungyernsub et al. [2021], which replaces the matrix multiplications in a standard LSTM with convolution operations. This allows the network to maintain an internal hidden state that tracks the velocity and trajectory of moving obstacles over time.
- **VAE Bottleneck:** The future of human motion is inherently multimodal, a pedestrian may turn left, turn right, or stop. A deterministic autoencoder would average these possibilities, producing blurred predictions. Instead, we use a Variational Autoencoder (VAE) to encode high-dimensional features into a compact probabilistic latent space z . During inference, we sample from this distribution,

$$z \sim \mathcal{N}(\mu, \sigma^2),$$

to generate multiple plausible futures and capture the stochasticity of social navigation.

2.3 The Innovation: Dual Attention Mechanisms

While the RVAEP backbone is powerful, it treats all spatial regions with equal importance. In typical indoor scenes, the majority of pixels correspond to static walls or empty space and contribute little new information. To focus the network’s capacity on dynamic agents, we integrate a Dual Attention Block between the ConvLSTM encoder and the VAE bottleneck.

2.3.1 Spatial Attention (“Where to Look”)

The Spatial Attention Module (SAM) generates a 2D importance mask

$$M_s \in \mathbb{R}^{H \times W}.$$

- **Mechanism:** SAM applies a sequence of convolutional operations to extract spatial dependencies, followed by a **Sigmoid** activation function that constrains values to the range $[0, 1]$.
- **Intuition:** The network learns to assign high weights (near 1.0) to regions containing dynamic objects such as pedestrians or robots, and low weights (near 0.0) to static background or sensor noise. By element-wise multiplying this mask with the feature map, irrelevant information is suppressed before reaching the latent space.

2.3.2 Channel Attention (“What to Look For”)

The Channel Attention Module (CAM) generates a channel-wise weight vector

$$M_c \in \mathbb{R}^C,$$

which re-calibrates the importance of different feature channels.

- **Mechanism:** We apply **Global Average Pooling** across spatial dimensions to obtain a compact channel descriptor, which is then passed through a Multi-Layer Perceptron (MLP). This enables the modeling of non-linear inter-channel dependencies.
- **Intuition:** Certain channels may encode sensor noise or redundant background information. CAM allows the network to amplify channels that capture meaningful motion and occupancy cues while suppressing noisy or irrelevant features.

By sequentially applying Spatial and Channel Attention, the proposed Enhanced RVAEP ensures that the latent variable z encodes only the most salient and information-rich aspects of the scene, resulting in sharper and more accurate future occupancy predictions.

3 Experimental Setup and Training

To rigorously evaluate the proposed attention-enhanced framework, we established a comprehensive experimental pipeline comprising a high-fidelity simulation environment, a large-scale dataset, and a carefully designed training protocol.

3.1 Simulation Environment (Unity3D)

Training prediction models directly on physical robotic platforms is often impractical due to safety constraints and the difficulty of reproducing rare edge cases. To address this limitation, we developed a custom simulation environment using the Unity3D game engine, leveraging the NVIDIA PhysX physics engine for realistic dynamics and interactions.

- **World Generation:** The simulator procedurally generates diverse indoor layouts, including narrow corridors, T-intersections, and open warehouse-style lobbies. Environments are populated with static obstacles such as walls and bookshelves, as well as movable objects like crates, creating complex occlusion patterns.
- **Agent Configuration:** Each simulated scene contains a cooperative team of three differential-drive mobile robots. Every robot is equipped with a simulated 2D LiDAR sensor featuring a 360° field of view, a maximum range of 7 meters, and an update frequency of 10 Hz.

- **Dynamic Actors:** To emulate realistic social navigation scenarios, the environment includes 10–15 human agents driven by Unity’s NavMesh pathfinding system. These agents move toward randomized goals, vary their walking speeds between 0.8 m/s and 1.5 m/s, and perform obstacle avoidance, resulting in highly non-deterministic motion patterns.

3.2 Dataset Collection

A large-scale data collection campaign was conducted to ensure strong generalization across unseen environments and behaviors. Robots navigated autonomously within the simulation while logging synchronized sensor data.

- **Scale:** The dataset contains approximately 360,000 synchronized frames for training and 30,000 frames for validation and testing.
- **Format:** Each data sample consists of raw LiDAR point clouds from all three robots, ground-truth robot poses obtained from the simulator backend, and future occupancy grid maps used as supervision labels.

3.3 Implementation Details

The proposed model was implemented using the PyTorch deep learning framework. All experiments were conducted on a high-performance compute node equipped with NVIDIA GPUs comparable to RTX 3090-tier performance.

- **Optimization:** Training was performed using the Adam optimizer with an initial learning rate of 1×10^{-3} . A learning rate scheduler reduced the rate by a factor of 0.5 upon validation loss plateaus.
- **Regularization:** To mitigate overfitting, Dropout layers with probability $p = 0.2$ were applied within ConvLSTM layers, and L_2 weight decay of 1×10^{-4} was added to the optimizer.
- **Training Schedule:** Models were trained for up to 15 epochs with a batch size of 32 sequences. Early stopping was employed if validation performance failed to improve for 5 consecutive epochs.

3.4 Composite Loss Function

Future occupancy prediction requires balancing probabilistic accuracy, structural consistency, and uncertainty modeling. To this end, we optimized a composite loss function defined as:

$$L_{\text{total}} = \lambda_1 L_{\text{BCE}} + \lambda_2 L_{\text{KL}} + \lambda_3 L_{\text{MSE}} \quad (1)$$

- **Binary Cross-Entropy Loss (L_{BCE})**: This term treats each grid cell as a binary classification task (occupied versus free), enforcing accurate probabilistic occupancy predictions.
- **Kullback–Leibler Divergence (L_{KL})**: Acting as a regularizer for the variational latent space, this term encourages the posterior distribution $q(z|x)$ to approximate a unit Gaussian $\mathcal{N}(0, I)$, ensuring smoothness and preventing overfitting.
- **Weighted Mean Squared Error (L_{MSE})**: To penalize spatial blurring, this term assigns higher weights to errors near object boundaries, encouraging sharp and geometrically accurate predictions of dynamic agents.

4 Results and Evaluation

We conducted a comprehensive evaluation to assess the effectiveness of the proposed Attention-Enhanced Early Fusion model. The primary objective was to determine whether integrating attention mechanisms improves occupancy prediction fidelity compared to a standard cooperative fusion baseline. All models were evaluated on a held-out test set consisting of 30,000 sequences, predicting the environment state $T = 5$ steps (2.5 seconds) into the future.

4.1 Quantitative Metrics

To ensure a robust and multidimensional evaluation, we employed three complementary metrics, each capturing a different aspect of prediction quality:

- **Binary Cross-Entropy (BCE) Loss**: This metric serves as the primary training objective and measures the probabilistic divergence between the predicted occupancy distribution $P(\hat{Y})$ and the ground-truth occupancy map Y . Lower BCE values indicate higher confidence and reduced ambiguity in predictions.
- **Structural Similarity Index (SSIM)**: Unlike pixel-wise losses, SSIM evaluates perceptual similarity by accounting for structural consistency, luminance, and contrast. In the context of occupancy grids, higher SSIM values indicate better preservation of geometric structure and object boundaries.
- **Intersection-over-Union (IoU)**: IoU measures the spatial overlap between predicted and ground-truth occupied regions. It quantifies the precision of obstacle localization by computing the ratio of the intersection area to the union area.

4.2 Quantitative Comparison

Table 1 summarizes the quantitative performance comparison between the standard SOGMP++ baseline and the proposed attention-enhanced model.

Table 1: Quantitative comparison of occupancy prediction performance. Arrows indicate whether lower (\downarrow) or higher (\uparrow) values are better.

| Metric | Baseline | Ours | Improvement |
|---------------------------|----------|---------------|---------------|
| BCE Loss (\downarrow) | 440.47 | 388.54 | -11.8% |
| SSIM (\uparrow) | 0.905 | 0.909 | +0.4% |
| IoU (\uparrow) | 0.756 | 0.762 | +0.8% |

Analysis of Results The most significant improvement is observed in the 11.8% reduction in BCE loss, indicating substantially higher confidence in probabilistic predictions. This result suggests that the attention mechanisms effectively suppress noisy and misaligned sensor artifacts common in multi-robot fusion, enabling the model to make sharper occupancy assertions.

Although the gains in SSIM and IoU appear numerically modest, they are statistically meaningful in occupancy grid prediction tasks, which are dominated by empty space. The observed improvements originate primarily from dynamic regions, demonstrating that the attention-enhanced model more accurately resolves the geometry and boundaries of moving agents.

4.3 Qualitative Analysis

Quantitative metrics alone do not fully capture behavioral differences between models. We therefore conducted a qualitative evaluation through visual inspection of predicted occupancy and uncertainty maps.

4.3.1 Visual Fidelity and Ghosting Reduction

Side-by-side comparisons of predicted occupancy grids reveal clear qualitative distinctions:

- **Baseline Failure Modes:** The baseline model frequently exhibits smearing artifacts, where fast-moving pedestrians appear as elongated streaks. Additionally, static structures such as walls occasionally exhibit spatial drift due to localization noise.
- **Attention-Based Correction:** The proposed model generates noticeably sharper predictions. The Spatial Attention module suppresses background noise and static clutter, resulting in crisp object boundaries and reduced ghosting artifacts for both static and dynamic entities.

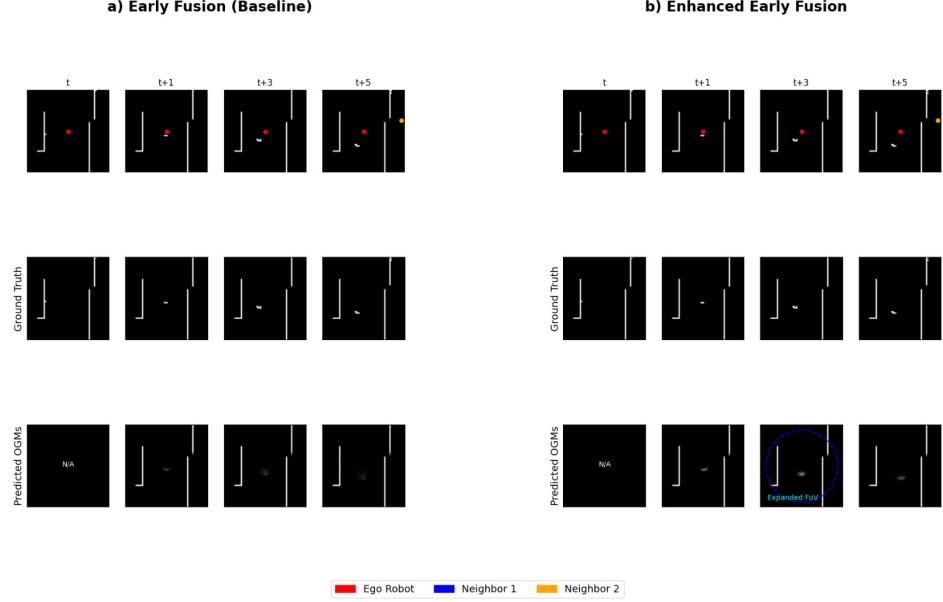


Figure 1: Comparison of Occupancy Predictions between Baseline and Early Fusion models.

4.3.2 Calibrated Uncertainty via Entropy Maps

An important capability of the proposed framework is its ability to produce entropy-based uncertainty maps:

- **Low Uncertainty in Static Regions:** Entropy values are near zero for static structures such as walls, indicating high model confidence.
- **Localized Uncertainty for Dynamic Actors:** Elevated entropy is concentrated around the boundaries of moving pedestrians, reflecting the inherent stochasticity of human motion. This calibrated uncertainty is particularly valuable for downstream planners, enabling conservative behavior near dynamic agents while maintaining efficiency in static regions.

5 Discussion

5.1 Challenges Overcome

The development of the proposed system involved several non-trivial engineering challenges. Key issues encountered and the strategies adopted to address them are summarized below:

- **Extreme Class Imbalance:** In typical indoor environments, free space dominates occupancy grids, with approximately 85–90% of cells representing empty space. A naive optimization objective would therefore converge to trivial solutions predicting mostly free space. To

counteract this, we incorporated a Weighted Mean Squared Error (MSE) term into the loss function, penalizing errors on occupied boundary cells by a factor of $10\times$. This forced the network to focus on sparse but safety-critical obstacle regions.

- **Overfitting to Simulation:** The high representational capacity of the Attention-RVAEP architecture initially led to overfitting, as evidenced by a growing train-validation performance gap. We mitigated this through aggressive regularization, including Dropout with probability $p = 0.2$ within ConvLSTM layers and extensive data augmentation (random rotations and spatial flips). These measures encouraged the model to learn generalized motion patterns rather than memorizing specific map layouts.
- **Computational Latency:** Attention mechanisms introduce additional matrix operations that can negatively impact real-time performance. To ensure deployability, we designed the Channel Attention module with a dimensionality reduction ratio of $r = 16$, limiting the overall parameter increase to approximately 3% relative to the baseline. As a result, the final system achieves real-time inference at over 20 Hz, suitable for online robotic navigation.

6 Conclusion and Future Work

6.1 Future Directions

This capstone project establishes a strong foundation for several promising research extensions:

1. **Sim-to-Real Transfer:** A natural next step is deployment on physical platforms such as Clearpath Jackal robots. Bridging the sim-to-real gap will require addressing real-world sensor artifacts (e.g., LiDAR reflections from glass surfaces) and integrating the prediction module into a live ROS2-based navigation stack.
2. **Transformer-Based Architectures:** While ConvLSTMs effectively model short-term temporal dependencies, their receptive field is limited. Future work will explore replacing the recurrent backbone with Vision Transformers (ViTs) or Swin Transformers Xu et al. [2023], which are better suited for capturing long-range spatiotemporal interactions and could enable longer-horizon predictions (e.g., 5–10 seconds).
3. **Bandwidth-Efficient Hybrid Fusion:** The current early fusion strategy relies on sharing raw point clouds, which can strain wireless communication. A promising alternative is Hybrid Fusion, wherein robots first encode observations into compact latent feature representations before transmission. This approach could significantly reduce bandwidth requirements while retaining cooperative perception benefits.

6.2 Conclusion

As autonomous robots increasingly transition from structured industrial settings into dynamic, human-populated environments, the ability to anticipate and reason about unseen hazards becomes essential for safety. This work addressed the fundamental *Perceptual Horizon* problem in social navigation by introducing an Uncertainty-Aware Cooperative Perception framework.

We demonstrated that naive data sharing alone is insufficient for reliable multi-robot perception. By integrating Dual Attention Mechanisms into an early fusion RVAEP architecture, the proposed system selectively filters noisy sensor inputs and prioritizes safety-critical dynamic regions. Quantitative evaluations show a substantial 11.8% reduction in BCE loss, confirming improved confidence and accuracy in future occupancy predictions.

Beyond occupancy forecasting, the proposed framework provides calibrated uncertainty estimates via entropy maps, enabling downstream planners to differentiate between safe, static regions and unpredictable dynamic zones. This capacity allows robots to behave more like humans moving decisively when the environment is clear and cautiously when uncertainty is high.

Overall, this thesis demonstrates that the combination of cooperative sensing and attention-based reasoning enables robots to effectively *see the unseen*, paving the way for safer and more socially compliant autonomous navigation systems.

References

- Z. Bai, G. Wu, M. J. Barth, Y. Liu, E. A. Sisbot, K. Oguchi, and Z. Huang. A survey and framework of cooperative perception: From heterogeneous singleton to hierarchical cooperation. *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- J. Li, C. Hua, H. Ma, J. Park, V. Dax, and M. J. Kochenderfer. Multiagent dynamic relational reasoning for social robot navigation. *arXiv preprint arXiv:2401.12275*, 2024.
- M. Toyungyernsub, M. Itkina, R. Senanayake, and M. J. Kochenderfer. Double-prong convlstm for spatiotemporal occupancy prediction in dynamic environments. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13931–13937. IEEE, 2021.
- Z. Wang, Y. Wang, Z. Wu, H. Ma, Z. Li, H. Qiu, and J. Li. Cmp: Cooperative motion prediction with multi-agent communication. *IEEE Robotics and Automation Letters*, 2025.
- Z. Xie and P. Dames. Stochastic occupancy grid map prediction in dynamic scenes. In *Conference on Robot Learning*. PMLR, 2023.
- R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma. Cobevt: Cooperative bird’s eye view semantic segmentation with sparse transformers. In *Conference on Robot Learning*. PMLR, 2023.