CS 7643 Milestone Report: Learning Costmap Generation from RGB & Depth for Mobile Robot Navigation

Rut Santana Ibrahim Alshayeb Vineet Kulkarni Meera Ranjan

Abstract

We aim to learn traversability costmaps directly from RGB + Depth images and assess them with both perception metrics and planner-level outcomes. Since the proposal, we implemented the end-to-end data pipeline (NYU and KITTI), dataset loaders, training/evaluation scripts, losses/metrics, and configuration files. This draft summarizes progress, clarifies hypotheses, outlines the methodology, and provides placeholders for preliminary results to be filled by teammates.

1 Introduction

Problem Motivation. Safe autonomous navigation requires reliable costmaps marking free space and obstacles for planners (e.g., A*, RRT*). Classical pipelines generated from depth/LiDAR can be brittle across domains and sensors. We study learning costmaps from RGB + Depth (RGBD) to improve robustness while preserving planner compatibility.

InputsOutputs Training Conditions. Input is a single RGBD image $I \in \mathbb{R}^{H \times W \times 4}$; output is a local, egocentric costmap $C \in [0, 1]^{64 \times 64}$. We also evaluate a binarized occupancy map by thresholding C. Supervision uses costmaps derived from depth-based heuristics. Optimization uses Adam with a composite objective of L1 and Dice losses (and an optional boundary-aware term).

Datasets and splits (paired .npz). NYU: 523 train, 131 val. KITTI: 438 train, 433 val. All pairs are standardized to 4-channel inputs and 64×64 targets.

Goal. Compare classical depth-to-cost mapping with UNet, ViT, and Hybrid CNN+Transformer models, measuring MAE, IoU, PrecisionRecallF1 on held-out splits and, later, planner performance. **Planner status:** Planner-in-the-loop evaluation with A* and RRT* is *not yet performed in this milestone* and is planned as future work.

2 Related Work

Our proposal emphasized learning traversability or BEV maps from vision and connecting predictions to planning utility. *TerrainNet* [3] fuses semantic and geometric cues for high-speed off-road traversability and stresses boundary fidelity and planning-aware metrics. A simplified U-Net for Mars rovers [4] highlights efficiency for resource-limited platforms. Transformer-based BEV mapping such as Trans4Map [5] improves global consistency; camera-only pipelines [6, 7] show promise without LiDAR but can be sensitive to depth errors and calibration. Preference-conditioned costmaps [8] offer flexibility that is complementary to our fixed traversability objective. We ground our models in foundational dense prediction with U-Net [1] and ViT [2]. Compared to prior work, our focus is a unified pipeline that standardizes supervision from depth-derived costmaps on indoor/outdoor datasets, compares UNet, ViT, and Hybrid encoders under identical training and metrics, and evaluates perception quality with a planned extension to planner-level outcomes (A*, RRT*).

3 Methodology (Tentative Technical Approach)

3.1 Data Processing and Labels

We implemented scripts to prepare NYU and KITTI, discover RGBDepth pairs, and generate targets. Depth is projectedprocessed into an egocentric grid and converted to a normalized costmap $C \in [0,1]^{64\times64}$. The same heuristic is used across datasets to ensure label consistency.

3.2 Models

We compare three families with a common lightweight decoder to a 1-channel output. The **UNet** baseline uses an encoder–decoder with skip connections. The **ViT** variant employs patch embedding, a transformer encoder, and a convolutional upsampling decoder. The **Hybrid** model combines a CNN stem with a transformer bottleneck followed by a CNN decoder to fuse local and global context. All models produce a 64×64 map (resize applied if needed).

3.3 Objective and Metrics

Training loss uses $\mathcal{L} = \lambda_{\ell_1}\,\mathcal{L}_{\ell_1} + \lambda_d\,\mathcal{L}_{\text{Dice}}(+\lambda_b\,\mathcal{L}_{\text{boundary}})$. Evaluation reports Mean Absolute Error (MAE) on continuous cost and IoU, Precision, Recall, and F1 on a binarized map (threshold τ=0.5 by default).

3.4 Evaluation Protocol

We train per-dataset (NYU, KITTI) and evaluate on held-out validation splits. Optional cross-domain tests (NYU→KITTI and vice versa) assess generalization. For this milestone, we report *perception metrics only*; planner-in-the-loop evaluation (A*, RRT*) of predicted costmaps is future work.

3.5 Hypotheses

We test three hypotheses. **H1 (Modality):** RGBD inputs outperform RGB-only for IoU and MAE when holding architecture and schedule fixed. **H2 (Architecture):** the Hybrid model achieves higher IoU than pure UNet or ViT at similar parameter budgets. **H3 (Objective):** combining L1 with Dice improves IoU over L1-only by emphasizing occupied regions and boundaries.

4 Baseline Results & Trials of Your Method

This section satisfies the milestone's requirement to report baseline and preliminary results. We trained/evaluated UNet, ViT, and Hybrid models on NYU and KITTI using identical splits, losses (L1+Dice), and threshold τ=0.5 for binarized metrics.

4.1 Baselines

We include a classical depth-to-cost baseline (thresholding, morphology, distance transform) and simple sanity checks (all-free and all-obstacle predictions) to contextualize metric behavior.

4.2 Quantitative Results

Table 1: NYU validation metrics. Threshold τ=0.5.

Method	MAE ↓	IoU ↑	Precision	Recall	F1	Params (M)
UNet	0.0139	0.9168	0.9526	0.9607	0.9566	4.2
ViT	0.0063	0.9750	0.9880	0.9860	0.9870	–
Hybrid	0.0088	0.9680	0.9830	0.9840	0.9840	–

Table 2: KITTI validation metrics. Threshold τ=0.5.

oprule Method	MAE ↓	IoU ↑	Precision	Recall	F1	Params (M)
UNet (KITTI only)	0.2083	0.4803	0.6148	0.6989	0.6384	4.2
UNet (NYU → KITTI TL)	0.2015	0.4999	0.6313	0.6956	**0.6514**	4.2
ViT	0.1890	0.4940	0.5990	**0.7230**	0.6500	–
Hybrid	**0.1740**	**0.5270**	**0.7090**	0.5850	0.6360	–

Summary. On NYU, *ViT* achieves the best overall performance across MAE/IoU/F1, with *Hybrid* close behind and *UNet* trailing by 5–6 IoU points. On KITTI, *Hybrid* attains the best MAE (0.174) and IoU (0.527) with strong precision, while recall is lower; *UNet* with transfer learning from NYU yields the best F1 (0.651) and improves over the KITTI-only UNet by +1.96 IoU points and +1.3 F1 points; *ViT* reaches the highest recall (0.723) and competitive F1 (0.650). The trends suggest indoor scenes (dense depth, regular geometry) favor global context from transformers, while outdoor KITTI benefits from Hybrid's local+global fusion and from cross-domain pretraining.

4.3 Implementation Notes

UNet. Architecture based on milesial/Pytorch-UNet with light edits. Training used Adam, L1+Dice loss, and a final `sigmoid` applied prior to L1/Dice and thresholding for metrics ($\tau = 0.5$). A transfer-learning run (pretrain on NYU, fine-tune on KITTI) improved KITTI metrics: IoU +1.96 pts (0.4803 → 0.4999), F1 +1.3 pts (0.6384 → 0.6514), and MAE from 0.2083 → 0.2015. Profiled model: **4.2M** params, **6.72 GFLOPs**, **1.75 ms/frame** on Tesla V100 (~571 FPS).

ViT. Patch embedding with transformer encoder and a convolutional upsampling decoder to a 1-channel output. Same optimizer and loss. ViT excelled on NYU (MAE 0.0063, IoU 0.975, F1 0.987) and achieved the highest recall on KITTI (0.723), indicating stronger sensitivity to thin/fragmented obstacles but with some over-segmentation (lower precision).

Hybrid. CNN stem with a transformer bottleneck and CNN decoder for local/global fusion. Heads match UNet/ViT for a fair comparison. Hybrid is competitive on NYU and leads MAE/IoU on KITTI (MAE 0.174, IoU 0.527) with high precision (0.709). Lower recall than ViT suggests potential benefit from boundary or focal terms and augmentation targeted at outdoor clutter.

4.4 Qualitative Results

Include side-by-side panels: RGB, Depth, baseline costmap, predicted costmap, and binarized occupancy overlay.

5 Next Steps

Near-term (1–2 weeks). Run a classical depth-to-cost baseline and add it to Tables 1–2; perform a hyperparameter sweep (lr, batch size, λ_d, data augmentations); and enable boundary or focal terms to raise KITTI recall for Hybrid without sacrificing precision.

Mid-term (3–4 weeks). Conduct ablations for H1–H3: modality (RGB vs RGBD), architecture (UNet vs ViT vs Hybrid at matched parameters), and objective (L1 vs L1+Dice vs +Boundary); run cross-domain tests (NYU→KITTI and KITTI→NYU) including transfer learning; and add calibration analysis (precision–recall across thresholds).

Endgame. Perform planner-in-the-loop evaluation (A* on grids, RRT* in continuous space) using predicted costmaps; report success, collisions, path cost, and planning time; finalize error analysis (thin obstacles, distant clutter), profile models (params, FLOPs, latency), and complete writing.

References (placeholder)

References

[1] Ronneberger, O., Fischer, P., Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI, 2015.

[2] Dosovitskiy, A., et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR, 2021.

[3] Meng, X., et al. TerrainNet: Visual Modeling of Complex Terrain for High-speed, Off-road Navigation. 2023.

[4] Qiu, R., and Lloyd, V. Reduced Image Classes in Modified U-Net for Mars Rover Navigation. 2025.

[5] Chen, C., et al. Trans4Map: Revisiting Holistic BEV Mapping from Egocentric Images with Vision Transformers. 2022.

[6] Bochare, A. Camera-Only Bird's Eye View Perception: A Neural Approach to LiDAR-Free Mapping for Autonomous Vehicles. 2025.

[7] Chang, et al. BEVMap: Map-Aware BEV Modeling for 3D Perception. 2024.

[8] Mao, L., et al. PACER: Preference-conditioned All-terrain CostMap Generation. 2025.

[9] Godard, C., Aodha, O. M., Brostow, G. Monodepth2. ICCV, 2019.

[10] Chen, C., et al. Deep Driving. ICCV, 2015.

[11] Tamar, A., et al. Value Iteration Networks. NeurIPS, 2016.

Appendix (Figures Only)

Figure 1: Qualitative examples (placeholder): RGB, Depth, classical baseline costmap, predicted costmap, binarized occupancy overlay.