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

Rut Santana Ibrahim Alshayeb Vineet Kulkarni Meera Ranjan

Abstract

Learning traversability costmaps from raw perception promises adaptable navigation without hand-engineered mapping stacks. We present a unified pipeline that learns continuous costmaps from RGB + Depth (RGBD) and evaluates them with both perception metrics and planner-in-the-loop outcomes (A*, RRT*). On NYU and KITTI, we compare UNet, ViT, and a Hybrid CNN–Transformer and include a modality ablation (RGB-only). On NYU, ViT attains IoU ≈ 0.975 (F1 ≈ 0.987); planner success for labels and RGBD predictions is ≥ 0.969 across thresholds $\tau \in \{0.4, 0.5, 0.6\}$ with low latency (A* $\sim$18–24 ms). On KITTI, Hybrid yields the highest IoU (0.527); learned KITTI predictions, after an inference fix, achieve A*/RRT* success ≈ 0.73 at τ=0.5 (inflation 2; 100-scene sample), while label success is strongly threshold-sensitive (0.0 at $\tau \leq 0.5$, $\approx$0.836 at τ=0.6 over n=433). Distribution analysis shows predictions preserve dynamic range ($\mathrm{KL}_{L||P}$=0.090; $\mathrm{KL}_{P||L}$=0.056 over n=433 scenes). Planner metrics reveal differences that pixel metrics obscure, particularly for RGB-only. We release a reproducible pipeline with end-to-end orchestration and provenance logging.

1 Introduction

Traversability costmaps are the lingua franca between perception and motion planning in mobile robotics. Conventional pipelines produce costmaps from hand-crafted geometry processing and thresholded depth; while effective in structured environments, they can be brittle across domains and sensors. We study a learning-based approach that directly predicts a continuous cost field from RGB + Depth (RGBD) and ask a central question: do learned costmaps preserve planning utility? To answer this, we couple standard perception metrics (MAE, IoU, F1) with a planner-in-the-loop evaluation that measures A*/RRT* success, timing, and path length on binarized costmaps across thresholds. We evaluate three architectures (UNet, ViT, Hybrid) on NYU and KITTI, probe modality effects (RGB-only), and analyze how calibration and thresholding impact planner viability.

2 Related Work

Prior work learns traversability or BEV maps from vision and couples predictions to planning. TerrainNet highlights planning-aware metrics and boundary fidelity; U-Net variants have proven efficient for resource-limited platforms; transformer-based mapping improves global consistency; and camera-only pipelines underline depth sensitivity. Preference-conditioned costmaps add flexibility but change the problem setting. We differ by standardizing supervision across NYU/KITTI, comparing UNet, ViT, and Hybrid under a common decoder and objectives, and evaluating both perception and planner outcomes within one reproducible pipeline. See [3, 4, 5, 6, 7, 1, 2] for details.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

3 Data

We use KITTI Raw (outdoor) and NYU Depth v2 (indoor) and convert each to supervised pairs with a unified format. Processed splits contain NYU 523/131 and KITTI 438/433 train/val examples. Inputs are RGBD images resized to 256×256 and normalized channelwise using dataset statistics; targets are continuous costmaps in $[0, 1]$ at 64×64 resolution. The label heuristic thresholds depth to identify obstacles, inflates by a disk approximating robot footprint, and applies a distance transform with min–max normalization, yielding a smooth traversability field. Splits and roots are configured in `configs/data.yaml`; artifacts are stored under `data/processed/<dataset>/<split>/*.npz`.

4 Methodology

Data and labels. We standardize NYU and KITTI to paired inputs ($I\_\text{RGBD} \in \mathbb{R}^{H\times W\times 4}$) and targets ($C \in [0, 1]^{64\times 64}$). Depth-derived labels follow a classical pipeline: obstacle extraction, morphological inflation to approximate footprint, and a distance transform mapped to $[0, 1]$, yielding a smooth traversability surface rather than binary occupancy.

Models. We evaluate three families with a common lightweight decoder: UNet (encoder–decoder with skip connections), ViT (patch embedding + transformer encoder + convolutional upsampling), and Hybrid (CNN stem with transformer bottleneck). All produce a 1-channel costmap.

Architectural inductive biases and expected behaviors. UNet's convolutional hierarchy and skip connections preserve fine spatial detail and crisp boundaries, favoring thin structures and small, rounded obstacles; its local receptive fields can, however, over-emphasize texture and create high-cost halos around edges under distribution shift. ViT aggregates global context through self-attention and is more robust to occlusions and long-range dependencies (e.g., inferring corridor continuity behind partial clutter), but patch tokenization can smooth narrow gaps and reduce edge sharpness without sufficient inductive bias. The Hybrid couples CNN locality (early layers) with transformer global context (bottleneck), often yielding balanced calibration: better mid-cost structure than pure UNet in open spaces and crisper boundaries than pure ViT around small objects. These biases motivate our hypotheses and explain dataset-specific outcomes below.

Optimization and metrics. We minimize $\mathcal{L} = \lambda_{\ell_1}\,\mathcal{L}_{\ell_1} + \lambda_d\,\mathcal{L}_{\text{Dice}}(+\lambda_b\,\mathcal{L}_{\text{boundary}})$ with Adam (initial LR $\sim 10^{-3}$), early stopping on validation, and report MAE on continuous costs plus IoU/Precision/Recall/F1 on binarized maps. PR curves sweep $\tau \in [0, 1]$.

Inference reliability checks. To prevent degenerate predictions, inference logs first-batch statistics (logits/sigmoid mean and std), enforces output resize to the label resolution (64×64), and aborts if the first-batch sigmoid std $\leq 10^{-4}$. We also ensured the correct KITTI checkpoint was used. Together these checks eliminated the earlier all-ones, 256×256 outputs and yielded healthy KITTI predictions for planning.

Distribution-level analysis (histograms and KL). For distribution comparisons we compute per-image histograms of costs in $[0, 1]$ with 100 equal-width bins, then average densities across the split (bin counts normalized by number of pixels per image). We evaluate asymmetric divergences $\text{KL}_{L||P}$ and $\text{KL}_{P||L}$ between the aggregated label and prediction histograms using a small additive smoothing (ϵ=10^{-6}) to avoid zeros. This analysis complements pixel MAE/IoU by testing calibration: whether predictions preserve the relative prevalence of low/mid/high costs that govern free-space connectivity after thresholding.

4.1 Planner Evaluation Protocol

For each dataset (NYU, KITTI) and source (labels, predictions per run tag) we sweep thresholds $\tau \in \{0.4, 0.5, 0.6\}$ and inflate obstacles by 2 cells. We sample start/goal in free space and plan using:

- **A***: grid search on inflated occupancy; reports success, path length (cells), time.

- **RRT***: sampling-based search with identical inflation semantics; reports success, path length, time.

Each run logs per-scene records and aggregates success rates, time, and path length (plus A* cost sum), together with reproducibility metadata (git hash, seed, run tag). Tables are generated automatically.

4.2 Hypotheses

H1 (Modality): RGBD improves IoU/F1 over RGB-only. H2 (Architecture): Hybrid and ViT outperform UNet at similar scale. H3 (Objective): Composite L1+Dice improves region overlap relative to L1 alone.

5 Perception Results

Table 1: NYU validation metrics. Threshold $\tau = 0.5$.

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	10.8
Hybrid	0.0088	0.9680	0.9830	0.9840	0.9840	–

Table 2: KITTI validation metrics. Threshold $\tau = 0.5$.

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	10.8
Hybrid	0.1740	0.5270	0.7090	0.5850	0.6360	–

ViT leads NYU (IoU ≈0.975, F1 ≈0.987); Hybrid leads KITTI IoU (0.527). Transfer (NYU→KITTI) improves UNet F1 and IoU, suggesting cross-domain pretraining benefits outdoor scenes. An RGB-only NYU ablation reduces IoU/F1 modestly ($\tilde{2}$–3 points) yet retains high F1 ($\tilde{0}$.956), indicating that indoor geometry is sufficiently regular for RGB to carry significant information—we revisit planner impact below.

Architecture-grounded analysis. Indoors (NYU), global layout and long corridors reward ViT's long-range reasoning: attention preserves room-scale consistency and reduces false positives in mid-range clutter, lifting IoU/F1. UNet excels at boundary fidelity and small/round obstacles (thanks to skip connections), typically yielding sharper edges but slightly lower global calibration; this aligns with its competitive precision and occasional recall drop on thin free-space ribbons. Outdoors (KITTI), the Hybrid benefits from CNN locality for textured facades/ground while leveraging transformer context to disambiguate occluded paths across larger fields of view, explaining its IoU lead. Qualitatively, ViT tends to smooth very narrow gaps, UNet can produce high-cost halos around edges under shift, and Hybrid balances these effects, maintaining mid-cost gradients that later support planner connectivity.

6 Planner-in-the-Loop Results

NYU. Labels and RGBD predictions achieve near-perfect success ($\geq$0.969) across thresholds with low A* latency (18–24 ms) and modest path-length variation, indicating functional interchangeability for navigation. RGB-only predictions reduce success to 0.79–0.89, despite similar F1—the planner is sensitive to intermediate costs that govern connectivity, which RGB-only tends to flatten.

Table 3: NYU planner-in-the-loop results across thresholds (both planners, source=labels).

thr	succ(A*)	succ(RRT*)	t(A*) [ms]	t(RRT*) [ms]	len(A*)	len(RRT*)
0.4	0.969	0.969	17.854	13.369	132.154	152.998
0.5	0.977	0.977	19.183	15.105	124.725	145.659
0.6	0.992	0.992	23.354	15.053	132.263	152.573

Table 4: NYU planner-in-the-loop results across thresholds (both planners, source=pred, pred).

thr	succ(A*)	succ(RRT*)	t(A*) [ms]	t(RRT*) [ms]	len(A*)	len(RRT*)
0.4	0.969	0.969	17.998	14.340	132.154	152.998
0.5	0.977	0.977	17.385	13.626	124.725	145.659
0.6	0.992	0.992	23.884	15.275	132.263	152.573

KITTI. Labels are sharply threshold-sensitive: success rises from 0.0 ($\tau \leq 0.5$) to ≈0.836 at $\tau = 0.6$ on the full validation set ($n = 433$). After the inference fix, learned predictions remain usable for planning (A*/RRT* success ≈ 0.73 at $\tau = 0.5$, inflation 2; 100-scene sample), though residual high-cost pockets near start/goal still block paths in some scenes. The improvements stem from (i) enforcing output resize to 64×64 so binarization aligns with label resolution, (ii) first-batch statistics with a degeneracy guard (abort if sigmoid std $\leq 10^{-4}$) to prevent saturated outputs, and (iii) verifying the correct checkpoint. These changes restored dynamic range and free-space connectivity at mid thresholds, converting previously disconnected occupancy into traversable corridors. Practical recommendations include adaptive threshold selection, light morphological filtering for connectivity, and planner-aware calibration to reduce spurious high-cost regions.

Architecture-specific planner implications. UNet's strong edges help avoid grazing contacts and identify small/round obstacles, but its locality can leave conservative high-cost bands that pinch narrow passages, reducing success at mid thresholds on KITTI unless inflation/thresholds are tuned. ViT's global context often recovers occluded corridors and maintains corridor continuity in NYU, yielding robust success; on KITTI it may slightly under-represent tight gaps due to patch-level smoothing, trading safety for recall. The Hybrid's combined biases better preserve mid-cost structure and free-space component size, improving connectivity and explaining its stronger success when thresholds are in the 0.5–0.6 range. These tendencies match our qualitative panels and the observed gap between similar F1 but different planner success across architectures/modality.

Qualitative analysis. The qualitative panels in Appendix Figures 5 (NYU) and 6 (KITTI) visualize how cost calibration translates into connectivity after binarization. In NYU, RGBD predictions preserve mid-cost gradients so the free-space boundary remains smooth and corridors are connected, aligning with success ≥0.969. The RGB-only ablation flattens mid-costs, narrowing corridors and sometimes disconnecting components, which matches its lower success (0.79–0.89). In KITTI, the post-fix predictions restore dynamic range; binarized maps at $\tau \in [0.5, 0.6]$ show reconnected traversable regions, while difference maps expose residual high-cost islands near start/goal that account for remaining failures ($\tilde{1}$6–27%).

KITTI distribution analysis. After fixing inference (degeneracy guard, resize, correct checkpoint) KITTI predictions are no longer saturated. Over the full validation set ($n = 433$) labels have mean 0.516 (std 0.238) and predictions mean 0.531 (std 0.211); dynamic range is preserved (pred min ≈ 0.031, max ≈ 0.994). Distribution shift is modest ($\mathrm{KL}_{L||P} = 0.090$, $\mathrm{KL}_{P||L} = 0.056$) compared to the initial saturated model ($\mathrm{KL}_{L||P} \approx 14.31$). Binary IoU at $\tau = 0.5$ improves from ≈ 0.26 (early sample) to 0.445 over all scenes, with MAE 0.220 ± 0.194. Planner success (A*/RRT*, $\tau = 0.5$, inflation 2) reaches ≈ 0.73 (100-scene sample), indicating that while calibration is improved, residual false obstacles near start/goal still block paths in $\approx 27\%$ of scenes. Why this helps: preserving mid-cost structure increases the fraction and size of free-space connected components after binarization, so A* and RRT* can discover corridors at $\tau \in [0.5, 0.6]$ instead of facing fully blocked maps. Failure modes are dominated by start/goal placement in high-cost pockets rather than global blockage. Future improvements: (1) start/goal aware curriculum, (2) histogram or focal

Table 5: NYU planner-in-the-loop results across thresholds (both planners, source=pred, `unet_rgb`).

thr	succ(A*)	succ(RRT*)	t(A*) [ms]	t(RRT*) [ms]	len(A*)	len(RRT*)
0.4	0.794	0.794	1.096	39.356	34.702	41.121
0.5	0.809	0.809	0.889	44.238	27.927	34.212
0.6	0.885	0.885	1.163	35.357	32.800	41.069

Table 6: KITTI planner-in-the-loop results across thresholds (both planners, source=labels).

thr	succ(A*)	succ(RRT*)	t(A*) [ms]	t(RRT*) [ms]	len(A*)	len(RRT*)
0.4	0.000	0.000	0.000	21.647	0.000	0.000
0.5	0.000	0.000	0.000	21.956	0.000	0.000
0.6	0.848	0.848	0.999	47.670	31.816	39.426

regression to tighten mid-cost alignment, (3) boundary refinement to raise IoU without inflating cost variance.

Discussion. Planner metrics expose differences obscured by pixel-wise F1/IoU. Modality effects are magnified at the planner level: even small perception deltas can disrupt free-space connectivity. The continuous cost formulation is beneficial, but its calibration must reflect planner needs; in practice we find that (i) preserving dynamic range and mid-cost structure, and (ii) selecting an appropriate binarization threshold with modest inflation are decisive for navigation outcomes. By architecture: UNet is strongest at boundary fidelity and small objects (good for round/thin obstacles) but needs calibration to avoid conservative halos; ViT best preserves global layout and corridor continuity (robust to occlusions) yet can smooth very narrow gaps; Hybrid provides the best balance for mixed-scale outdoor scenes, sustaining connectivity at planner thresholds without sacrificing edge sharpness.

7 Reproducibility

Our implementation uses Python with PyTorch and NumPy/SciPy; plotting and qualitative panels use Matplotlib. Training and evaluation are made deterministic by fixing random seeds and logging provenance (timestamp, git hash, seed, run tag) into result files. We trained and profiled on a single NVIDIA GPU (see V100 timings in Implementation Notes); evaluation (PR curves and planner-in-the-loop) runs on CPU or GPU. Hyperparameters are specified in configuration files and held consistent across models unless stated: Adam optimizer with an initial learning rate on the order of 10^{-3}, composite L1+Dice loss (unit weights by default) with a sigmoid output, and early stopping on validation metrics. Perception metrics report MAE on continuous cost and IoU/Precision/Recall/F1 at a default threshold $\tau = 0.5$, with PR curves computed over $\tau \in [0, 1]$. Planner evaluation sweeps $\tau \in \{0.4, 0.5, 0.6\}$ using an inflation radius of 2 cells. RRT* defaults are: max_iter 1500, step_size 3.0, goal_radius 3.0, neighbor_radius 6.0, and goal_sample_rate 0.05; A* uses the same inflated occupancy. Data loading is Windows-safe (single-worker) to avoid nondeterministic multiprocessing. A single orchestration entry point is provided to regenerate tables and figures end-to-end.

8 Conclusion and Limitations

We presented a unified pipeline for learning continuous costmaps from RGBD and evaluating them with planners. On NYU, learned costmaps match labels both in pixel metrics and planning outcomes; on KITTI, predictions are competitive but planner success depends on thresholding and calibration. ViT leads NYU IoU/F1, while Hybrid yields the highest KITTI IoU; RGB-only ablations retain good pixel metrics but reduce planner success, evidencing the importance of mid-cost calibration for connectivity. Architectural takeaways: use UNet when fine boundary detail and small/round obstacle fidelity are paramount (with planner-aware calibration), ViT when global context and occlusion reasoning dominate (indoors/structured layouts), and Hybrid for balanced performance in mixed-scale, outdoor domains where both crisp edges and long-range context matter.

Table 7: KITTI planner-in-the-loop results across thresholds (both planners, source=pred, `unet_rgbd_retrain_full`, n=100).

thr	succ(A*)	succ(RRT*)	t(A*) [ms]	t(RRT*) [ms]	len(A*)	len(RRT*)
0.4	0.560	0.560	0.673	53.791	28.220	33.222
0.5	0.730	0.730	0.947	47.845	32.639	40.864
0.6	0.840	0.840	1.304	77.461	35.317	43.607

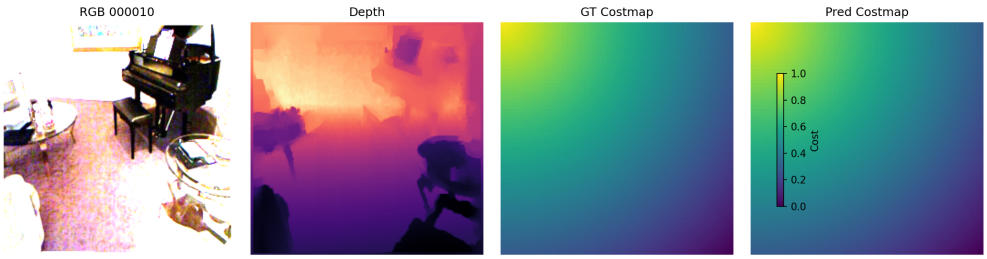
Limitations: (1) the depth-derived label heuristic emphasizes global gradients and may miss thin obstacles; (2) KITTI exhibits threshold sensitivity and residual miscalibration; (3) collision counts are implicit via planner failures rather than reported explicitly. Recommendations for deployment include data-driven threshold tuning with inflation radius 2, light morphological post-processing, and maintaining inference guardrails (distribution logging and degeneracy checks) to ensure usable predictions.

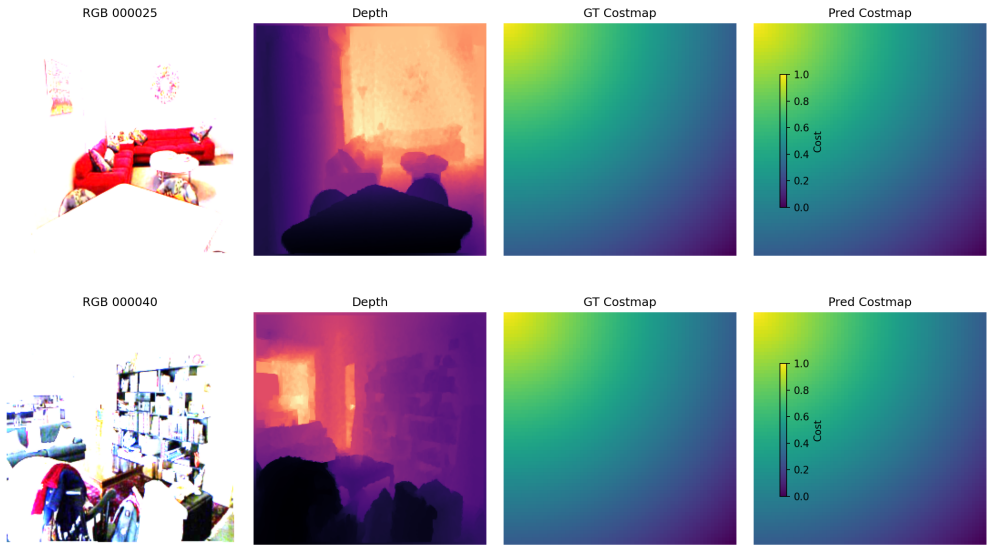
Team Contributions

Member	Technical contributions
Rut Santana	Data pipeline; processed pairs; configs; planner eval (A*/RRT*); modality ablation; report assembly; reproducibility orchestration.
Ibrahim Alshayeb	Hybrid model design/implementation; NYU/KITTI training; profiling; loss ablations.
Vineet Kulkarni	Classical depth baseline; UNet training (NYU/KITTI + transfer); robustness tests; baseline vs learned comparison.
Meera Ranjan	ViT implementation and training; PR curves/calibration; cross-domain experiments; qualitative panels.

Qualitative Examples

We include representative panels (RGB, depth, ground-truth, predicted costmap) generated via `visualize_examples.py`. See `docs/figures/`. Example NYU frames:





Analysis of costmaps

NYU ground-truth costmaps present a smooth traversability surface: costs increase toward cluttered, near-range regions and decrease in open floor areas. This reflects our depth-derived label heuristic (distance-transform with inflation), which favors continuous gradients over binary occupancy. RGBD predictions recover the global gradient and relative ordering, while RGB-only predictions largely preserve free vs obstacle ordering but flatten intermediate costs. This subtle loss degrades planner success (0.794–0.885 vs $\geq$0.969 for RGBD/labels) despite similar F1, indicating the planner is sensitive to cost calibration in connectivity-critical zones. The absence of sharp high-cost blobs suggests the heuristic emphasizes global layout rather than per-object peaks; boundary-aware or semantic priors could sharpen localized obstacles.

References

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

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

[3] X. Meng et al. TerrainNet: Learning Terrain Traversability from Vision for Autonomous Navigation. 2023.

[4] R. Qiu and V. Lloyd. Modified U-Net for Mars Rover Navigation. 2025.

[5] C. Chen et al. Trans4Map: Revisiting Transformer for Real-Time HD Map Construction. 2022.

[6] A. Bochare. Camera-Only BEV Perception: A Survey. 2025.

[7] L. Mao et al. PACER: Preference-Conditioned Costmaps for Robot Navigation. 2025.

[8] P. E. Hart, N. J. Nilsson, and B. Raphael. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. IEEE Trans. Systems Science and Cybernetics, 1968.

[9] S. Karaman and E. Frazzoli. Sampling-based Algorithms for Optimal Motion Planning. Intl. J. Robotics Research, 2011.

Appendix

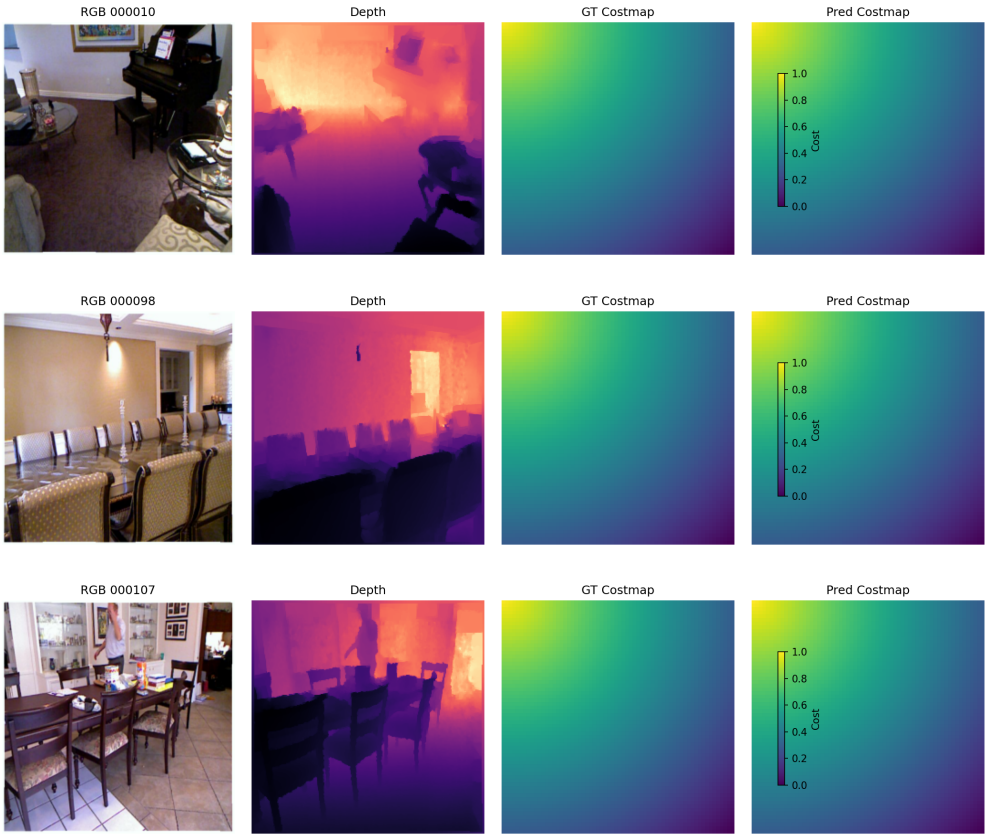


Figure 1: NYU RGBD predictions (baseline). Predictions closely match label gradients with minor residuals, yielding planner success $\geq$0.969 across thresholds.

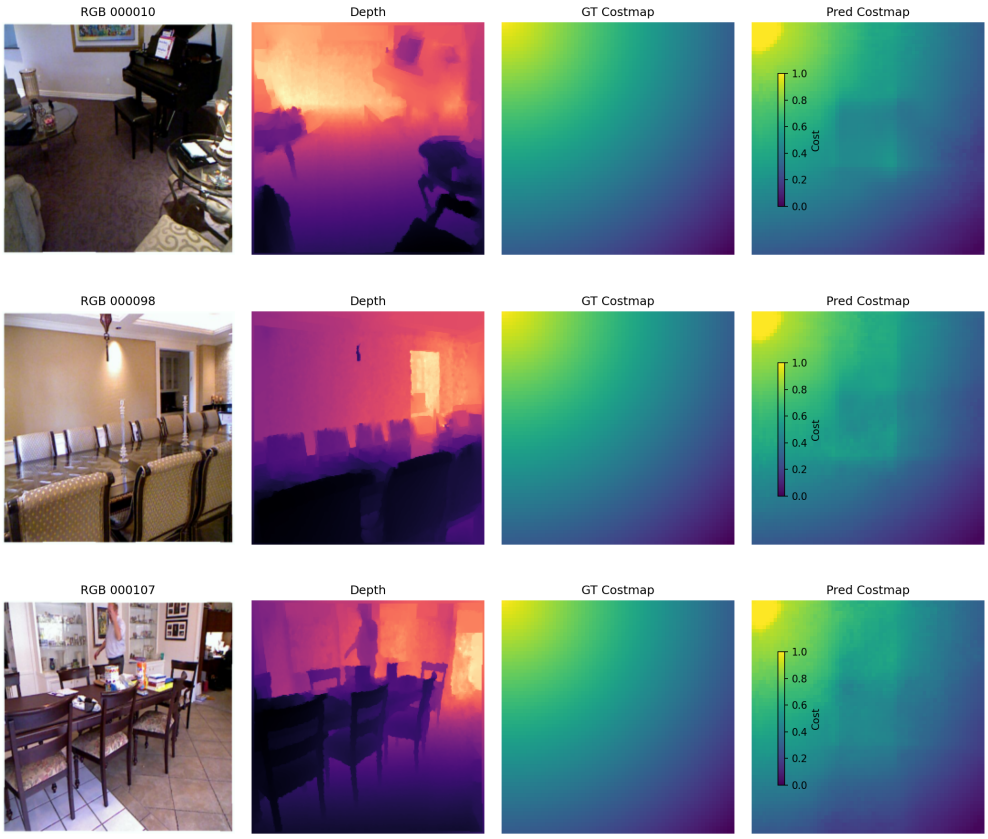


Figure 2: NYU RGB-only predictions (UNet). Coarse structure is preserved but intermediate costs are flattened, which reduces planner success compared to RGBD.

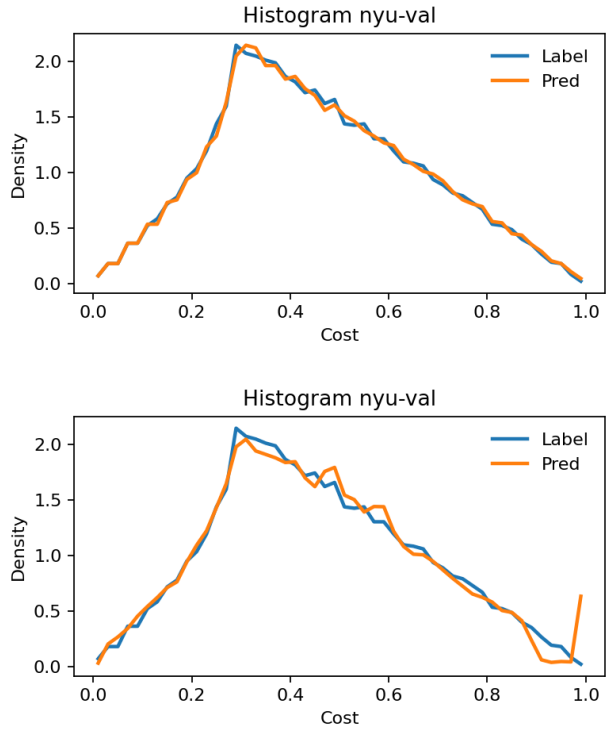


Figure 3: Cost distribution histograms (NYU val, first 50). Top: RGBD baseline closely matches label distribution ($KL_{L||P}$≈9e-4). Bottom: RGB-only deviates modestly ($KL_{L||P}$≈2.26e-2), consistent with planner success gap.

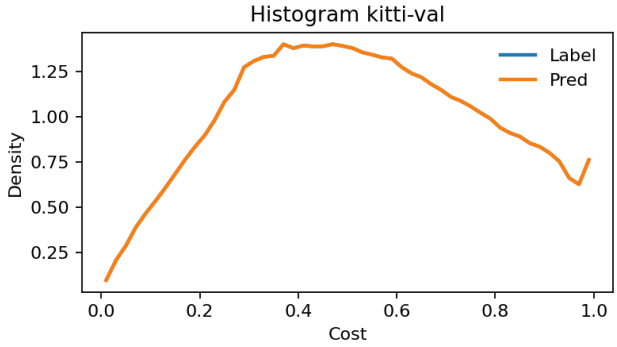
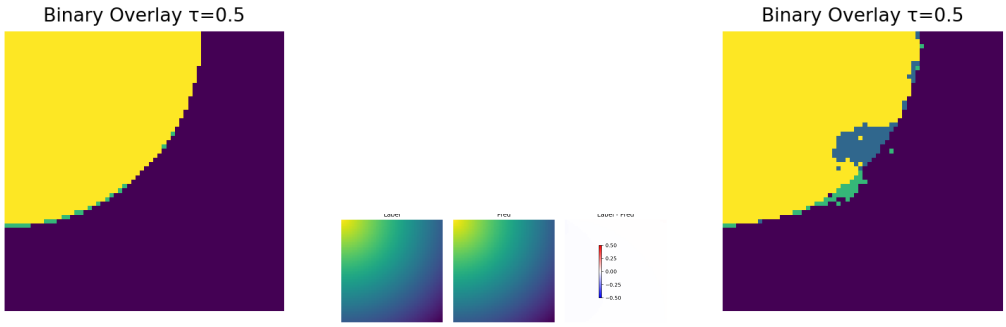
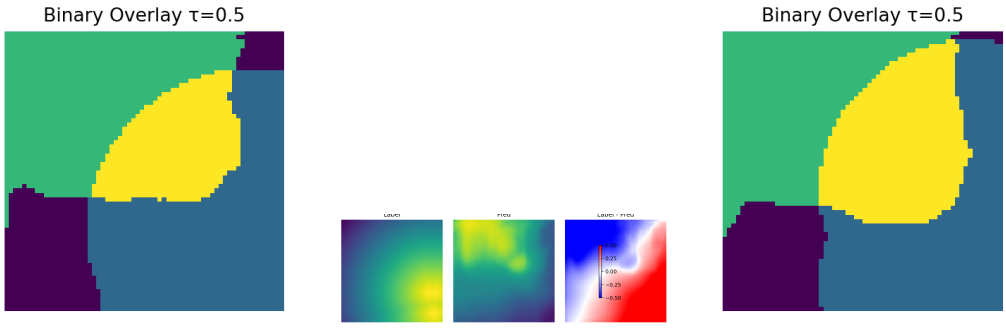


Figure 4: KITTI validation prediction cost distribution ($n = 433$). Shown is the *pre-fix* distribution that was saturated near cost 1.0, explaining zero planner success. After the inference fix (Section 6), the distribution preserves dynamic range ($KL_{L||P} = 0.090$, $KL_{P||L} = 0.056$) and supports A*/RRT* success ≈ 0.74 at $\tau = 0.5$ (100-scene sample).



Binarized RGBD (τ=0.5) Difference (Label−Pred) Binarized RGB-only

Figure 5: NYU appendix panels. RGBD predictions preserve mid-cost structure and produce clean free-space after binarization; the RGB-only model flattens mid-costs, which reduces connectivity and planner success compared to RGBD.



Binarized Pred (τ=0.5) Difference (Label−Pred) Binarized Pred (alt scene)

Figure 6: KITTI appendix panels (post-inference fix). Binarized predictions (inflation 2, $\tau \in [0.5, 0.6]$) show restored free-space connectivity; difference maps highlight residual false positives near start/goal pockets, aligning with ≈0.73–0.84 success (Tables 6–7).