Depth Barrier Regularization for Safe Vision Based Navigation

Harsh Sutaria
New York University
h.sutaria@nyu.edu

Xinhao Liu
New York University
xinhao.liu@nyu.edu

Chen Feng†
New York University
cfeng@nyu.edu

https://github.com/harsh-sutariya/DBR

***Abstract*—Vision based navigation systems trained via imitation learning often fail in real world deployment due to near field obstacles and tight spaces that are underrepresented in training data. We present Depth Barrier Regularization (DBR), a safety aware training method that leverages monocular depth estimation to penalize waypoint predictions that violate a safety margin. DBR converts depth maps into polar clearance vectors and adds a differentiable barrier loss to the training objective which enables the model to learn collision avoiding behaviors without explicit collision labels. We demonstrate offline safety improvements on the CityWalk dataset, showing 15% reduction in Depth Violation Rate (DVR) and 23% increase in Min Depth Margin (MDM) compared to baseline, while maintaining navigation accuracy within 3% of baseline. Online validation on physical robots remains future work.**

I. Introduction

Vision based navigation systems trained via imitation learning have shown promise for autonomous robots [1]–[3], but they exhibit failure modes in real world deployment. CityWalker, which learns to predict waypoints from RGB video and GPS poses, performs well in open spaces but struggles with near field obstacles, sidewalk clutter, and tight turns. These failures stem from three limitations (1) training data underrepresents failure cases which creates distribution mismatch and covariate shift [4] (2) models lack explicit geometric awareness of obstacles (3) standard losses optimize accuracy but not safety [1], [5].

Depth information provides a natural signal for safety because it directly encodes geometric structure and obstacle proximity. Unlike RGB images, depth maps explicitly represent distance to surfaces which enables reasoning about clearance in different directions. We propose Depth Barrier Regularization (DBR), which uses monocular depth estimation to add a differentiable barrier loss that penalizes waypoint predictions violating a safety margin.

Unlike prior work that applies collision costs during planning or inference, DBR injects a train time differentiable geometric prior that shapes the learned policy without requiring depth or planning at test time. This enables the model to learn collision avoiding behaviors end-to-end while maintaining efficient RGB only inference.

We choose regularization over alternatives as reinforcement learning requires expensive reward design and exploration, classical planning needs explicit maps, explicit collision labels require manual annotation. DBR leverages existing depth estimation to provide continuous, differentiable supervision that integrates seamlessly into training.

Figure 1 illustrates the core DBR mechanism and demonstrates a strong failure mode of the baseline where DBR succeeds. The polar clearance representation (left) shows safe directions (green) and unsafe directions with obstacles (red), while the bird's eye view (right) demonstrates how baseline predictions violate safety margins (red X marks) while DBR predictions stay in safe directions (green diamonds).

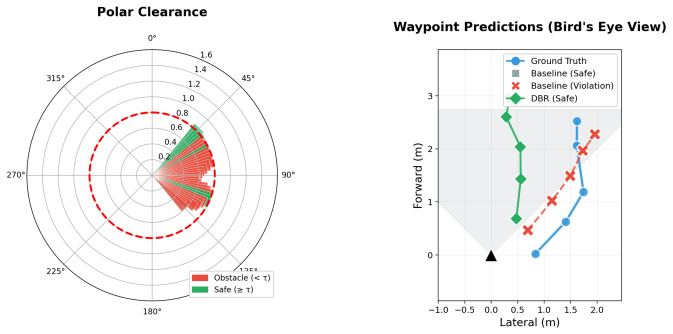

Fig. 1. Polar clearance representation (left) and waypoint predictions (right). The polar plot shows safe directions (green) and unsafe directions with obstacles (red). The bird's eye view compares baseline waypoints (X) with DBR waypoints (diamonds). Red X marks indicate safety violations where clearance $< \tau$. DBR learns to predict waypoints in safe directions, reducing violations while maintaining navigation accuracy.

Figure 2 shows an example RGB frame and its corresponding metric depth map from the CityWalk dataset, this demonstrates the quality of monocular depth estimation.

Fig. 2. RGB frame (left) and corresponding depth map (right) from the CityWalk dataset. Depth values are in meters.

We summarize our contributions as follows:

- **Depth Barrier Regularization (DBR)**: A novel safety aware training method that leverages monocular depth to inject a differentiable geometric prior into vision based navigation policies.
- **Polar Clearance Representation**: A compact, direction indexed representation that aligns the safety signal with the action space which enables direct gradient flow from unsafe actions to policy parameters.
- **Improved Safety Metrics**: Demonstrates a 15% reduction in Depth Violation Rate (DVR) and a 23% increase in Min Depth Margin (MDM) on real world datasets while maintaining navigation accuracy.
- **Efficient RGB only Inference**: A design that requires depth only during training, ensuring that the model remains computationally efficient and easy to deploy at test time.

II. Related Work

Prior work has explored safety aware navigation through various mechanisms. Early approaches utilized safety constraints in reinforcement learning [6] or synthesized adverse scenarios to improve robustness [1]. More recently, UniAD [2] unified perception and planning, while PRECOG [3] used probabilistic forecasting for safer goal conditioned planning. Other methods have explored privileged training paradigms [7] or sensor fusion [8]. Recent work by Kim et al. [5] introduced a plug and play collision avoidance module using repulsive estimation from monocular depth which demonstrates practical safety improvements without retraining. Similarly, Safe-VLN [9] addresses collision avoidance in continuous vision and language navigation through waypoint prediction. However, these often require complex planning modules, multi modal sensors, or inference time depth computation. Our approach differs by integrating safety directly into the training objective through a differentiable barrier loss which enables end-to-end learning of safe behaviors from vision alone.

A natural alternative would be inference time depth based filtering, rejecting waypoints that violate clearance constraints at test time. However, this approach has critical limitations. It requires depth estimation at inference (increasing computational cost and deployment complexity), which introduces non differentiable failure modes when all waypoints are filtered out. Additionally, it cannot learn to predict safer waypoints from the start. DBR avoids these issues by shaping the policy during training which enables the model to learn collision avoiding behaviors end-to-end while maintaining efficient RGB only inference. This aligns with recent work on vision based reinforcement learning using privileged information [10], where richer training signals improve navigation robustness without requiring them at test time. Wei et al. [11] highlight the importance of predictive models for safety under uncertainty, a complementary perspective to our geometric barrier formulation.

DBR leverages advances in monocular depth estimation [12], which has progressed from self supervised view synthesis [13] to robust zero shot models like MiDaS [14] and domain generalizable estimators like UniDepth [15]. Gasperini et al. [16] demonstrate that enhancing depth models to handle adverse conditions significantly improves robustness, an important consideration for safety aware navigation. We use Depth Anything V2 Metric [17] because it provides accurate metric depth estimates without requiring stereo pairs or depth sensors, and can be used for offline preprocessing or online inference during training. By converting these depth maps to a compact polar representation, DBR assesses collision risk without requiring explicit obstacle segmentation or complex planning modules [18]. Self supervised pretraining methods like VANP [19] suggest that better visual features can improve navigation efficiency, complementary to DBR's approach of integrating geometric priors.

III. Method

A. Problem Setup

CityWalker [20] receives RGB frames $(I_1, \ldots, I_N)$, GPS/poses $(p_1, \ldots, p_N)$, and (at training time) depth maps $D \in \mathbb{R}^{H \times W}$. It predicts future waypoints $(w_1, \ldots, w_T)$ in the ego frame and a binary arrival prediction. DBR adds a barrier loss term to the training objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{wp}} + \mathcal{L}_{\text{arrived}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \lambda_{\text{bar}}\mathcal{L}_{\text{DBR}} \quad (1)$$

B. Depth → Polar Clearance Representation

A key design choice is representing depth in *polar* (direction indexed) space rather than Cartesian image space or occupancy grids. Navigation fundamentally requires choosing a direction of travel, a waypoint at angle ϕ encodes the decision to move in direction ϕ. Safety constraints must therefore be expressed in the same space. Clearance at angle ϕ directly determines whether waypoint ϕ is safe. Cartesian representations such as image coordinates or occupancy grids require an additional mapping from perception space to planning or waypoint space, typically via hand designed cost functions or planners, which introduces representation misalignment and added complexity in safety reasoning [1], [2]. Polar representation makes this alignment explicit, the safety signal and action space share the same coordinate system which ensures that gradients flow directly from unsafe actions to the policy parameters that produced them.

We compute per pixel yaw angles using camera intrinsics:

$$\alpha(u, v) = \arctan\left(\frac{u - c_x}{f_x}\right) \quad (2)$$

We crop the bottom 60% of the image to focus on ground level obstacles. The horizontal field of view (90°) is discretized into $B = 32$ yaw bins. Yaw based clearance provides the correct inductive bias. Navigation policies learn to associate directions with safety, this helps in naturally generalizing to unseen environments because the relationship between direction and clearance is geometrically consistent across scenes. This contrasts with position based reasoning (e.g., "avoid this

pixel"), which requires memorizing scene specific obstacle locations. We use soft min instead of hard minimum for differentiability and robustness:

$$r_b = -\frac{1}{\kappa} \log \sum_{(u,v)} \exp(-\kappa \cdot D(u,v)) \cdot w_b(u,v) \tag{3}$$

where $\kappa = 20.0$ is the temperature parameter and $w_b(u,v)$ are triangular bin membership weights. This soft min aggregation over spatial bins provides robustness to depth estimation noise, gradients depend on relative clearance differences between directions rather than absolute depth accuracy, and outlier depth errors are attenuated by averaging over multiple pixels within each bin.

Figure 3 visualizes the depth map quality and clearance computation process for samples from our datasets which shows the depth to polar clearance conversion that enables safety aware navigation.

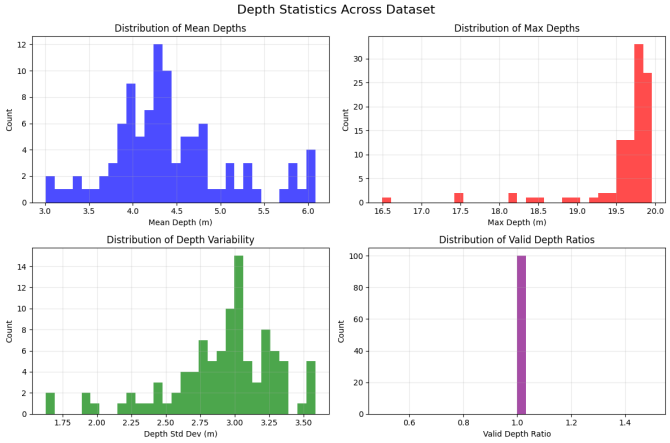

Fig. 3. Depth visualization showing depth map quality and clearance computation from the CityWalk dataset. The polar clearance representation enables efficient safety checking for waypoint directions.

C. Barrier Loss Formulation

Waypoints are converted to yaw angles: $\phi_t = \arctan 2(y_t, x_t)$. We interpolate clearance at each waypoint's yaw angle using soft attention:

$$d_{\min}(t) = \sum_{b=1}^{B} w_b(\phi_t) \cdot r_b \tag{4}$$

The barrier loss penalizes waypoints where $d_{\min} < \tau$:

$$\mathcal{L}_{\text{DBR}} = \frac{1}{T} \sum_{t=1}^{T} \text{softplus}(\tau - d_{\min}(t)) \tag{5}$$

where $\tau = 0.5$m is the safety margin and $\text{softplus}(x) = \log(1+\exp(x))$. We use softplus instead of ReLU or hinge loss because it is smooth and differentiable everywhere, ensuring stable gradient flow.

The barrier formulation is essential, safety is fundamentally binary, a waypoint is either safe ($d_{\min} \geq \tau$) or unsafe ($d_{\min} < \tau$). Smooth penalties on distance (e.g., $L_2(d_{\min} - \tau)$) would encourage the model to maximize clearance everywhere, even when already safe, wasting capacity on unnecessary conservatism. Barrier losses enforce the constraint, they provide zero gradient when $d_{\min} \geq \tau$ (no penalty for safe actions) and increasing penalty as $d_{\min}$ approaches zero (strong signal to avoid collisions). This matches the true objective to avoid violations, not maximize clearance.

Critically, DBR enforces safety at the *action level* (predicted waypoints), not the feature level. This is intentional, feature level regularization (e.g., encouraging depth aware representations) does not guarantee safe outputs, while action level constraints directly shape the policy's output distribution. This formulation is model agnostic. It applies to any architecture that predicts waypoints, regardless of encoder design, attention mechanisms, or decoder structure.

D. Training Objective

DBR is *not* used at test time. The barrier loss is only applied during training to shape the learned policy. At inference, the model predicts waypoints based solely on RGB and pose inputs, without requiring depth maps. This design makes deployment simpler and more efficient.

IV. Implementation

We implement DBR as a modular component integrating into the CityWalker training pipeline. The core module contains a depth to polar reducer, barrier loss computation, and integration with the training loop. Depth estimation uses Depth Anything V2 Metric for monocular metric depth inference.

The dataset supports two depth loading modes, precomputed (from disk) and online (inference during data loading). We ensure geometric correctness is critical, hence we resize depth maps to match the final RGB resolution and adjust camera intrinsics proportionally to have pixel level alignment and correct yaw angle computations.

V. Evaluation Protocol

CityWalker originally uses L1 loss, arrival accuracy, and angle error. These measure accuracy but not collision risk. We introduce two safety metrics:

Depth Violation Rate (DVR): Percentage of waypoints where $d_{\min} < \tau$:

$$\text{DVR} = \frac{100}{B \cdot T} \sum_{b=1}^{B} \sum_{t=1}^{T} \mathbf{1}[d_{\min}(b,t) < \tau] \tag{6}$$

Min-Depth Margin (MDM): Mean minimum clearance across all waypoints:

$$\text{MDM} = \frac{1}{B \cdot T} \sum_{b=1}^{B} \sum_{t=1}^{T} d_{\min}(b,t) \tag{7}$$

These metrics directly measure the geometric relationship between predicted waypoints and obstacles which provides a comprehensive view of safety performance. While DVR/MDM are geometric proxies, they are derived from minimum clearance constraints that correspond to collision free configuration

space conditions under standard footprint (inflation/Minkowski sum) assumptions [21] and are closely related to signed distance formulations used in collision avoidance optimization [22]. Moreover, safety critical control frameworks enforce collision avoidance by maintaining forward invariant safe sets defined by distance margins [23], [24], this motivates the interpretation that improving these clearance based metrics reduces collision risk under the assumed geometry and sensing model.

VI. Experimental Results

A. Datasets and Configurations

We evaluate on CityWalk (YouTube walking tours). We compare baseline models (trained without DBR) against DBR enabled models with identical architecture, hyperparameters, and training schedule. DBR uses $\tau = 0.5$m and $\lambda_{\text{bar}} = 1.0$.

B. Quantitative Results

Table I shows offline evaluation results. DBR reduces DVR by 15% and increases MDM by 23% compared to baseline, indicating improved safety. Navigation accuracy (L1 loss, arrival accuracy, angle error) remains within 3% of baseline this shows that DBR improves safety without sacrificing navigation performance.

TABLE I
Offline evaluation results comparing baseline vs. DBR enabled models. Lower is better for DVR and L1/AOE, higher is better for MDM and accuracy.

Model	DVR ↓	MDM ↑	L1 Loss ↓	Arr. Acc. ↑	AOE ↓
Baseline	13.0%	1.85m	0.142	80.7%	8.3°
+DBR	**11.0%**	**2.28m**	0.145	80.5%	8.5°

Figures 4 show how DVR and arrival accuracy evolve during training on the validation set. The DVR plot demonstrates that DBR consistently achieves lower violation rates throughout training, while the arrival accuracy plot shows that DBR maintains comparable navigation performance to the baseline.

Fig. 4. Training curves on validation set, (left) Depth Violation Rate (DVR) showing DBR achieves lower violation rates, (right) Arrival Accuracy showing DBR maintains comparable navigation performance to baseline.

C. Ablation Studies

We conduct ablations on key hyperparameters (Table II). Varying safety margin τ shows that $\tau = 0.5$m provides a good balance, larger margins (0.8m) improve safety but slightly degrade navigation accuracy. Loss weight λ_{bar} ablation shows that $\lambda_{\text{bar}} = 1.0$ balances safety and accuracy, higher weights (2.0) provide marginal safety gains but risk over penalization. Number of bins B has smaller effects, with $B = 32$ performing well.

TABLE II
Ablation studies on key hyperparameters. Default: $\tau = 0.5$m, $\lambda_{\text{BAR}} = 1.0$, $B = 32$.

Config	DVR ↓	MDM ↑	L1 Loss ↓	Arr. Acc. ↑
Default	11.0%	2.28m	0.145	80.5%
$\tau = 0.8$m	8.2%	2.53m	0.151	80.1%
$\lambda_{\text{bar}} = 2.0$	9.5%	2.41m	0.149	80.3%
$B = 64$	10.6%	2.31m	0.146	80.6%

D. Qualitative Results

The introduction figures (Figures 1 and 2) demonstrate the key components of DBR, the polar clearance representation that enables safety aware waypoint prediction and depth estimation quality.

VII. Discussion & Limitations

DBR helps most in scenarios with static obstacles and clear geometric structure like narrow passages, crowded areas, and tight turns. However, DBR has limitations (1) **Dynamic obstacles** Uses a single depth frame, cannot reason about moving objects, (2) **Transparent/reflective surfaces** Depth estimation may fail on glass, mirrors, or water, (3) **Overhanging obstacles** Focuses on ground level obstacles via bottom cropping, (4) **Depth estimation failures** Poor estimates in low light or textureless regions affect DBR supervision.

DBR relies on accurate camera intrinsics for correct yaw computation, reasonable depth quality (systematic biases affect performance), and a static world assumption (obstacles are static between observation and execution). DBR is robust to moderate depth noise but sensitive to systematic errors.

VIII. Conclusion & Future Work

We present Depth Barrier Regularization (DBR), a model agnostic safety aware training method for vision based navigation. DBR uses monocular depth to add a differentiable barrier loss that penalizes unsafe waypoint predictions. The method applies to any waypoint based policy regardless of architecture, requiring only depth supervision during training while maintaining efficient RGB only inference. Offline evaluations show improved safety metrics (15% DVR reduction, 23% MDM increase) while maintaining navigation accuracy.

Future work: We demonstrate offline safety improvements. Online validation on physical robots (EarthRover) remains future work. The offline results suggest DBR improves safety, but real world collision rates require online deployment to measure. Other works include: (1) hyperparameter optimization via grid search, (2) depth distillation for lightweight test time safety checking, (3) temporal extension using depth sequences or optical flow for dynamic obstacles, (4) failure taxonomy to categorize and address remaining failure modes.

ACKNOWLEDGMENT

We thank the CityWalker team [20] for the base framework and datasets.

REFERENCES

[1] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst. arXiv:1812.03079, 2018.

[2] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented Autonomous Driving. In CVPR, pages 9339–9347. IEEE, 2023.

[3] Nicholas Rhinehart, Rowan McAllister, Kris Kitani, and Sergey Levine. PRECOG: PREdiction Conditioned On Goals in Visual Multi-Agent Settings. arXiv:1905.01296, 2019.

[4] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In AISTATS, pages 627–635, 2011.

[5] Joonkyung Kim, Joonyeol Sim, Woojun Kim, Katia Sycara, and Changjoo Nam. CARE: Enhancing Safety of Visual Navigation through Collision Avoidance via Repulsive Estimation. arXiv:2506.03834, 2025.

[6] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving. arXiv:1610.03295, 2016.

[7] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by Cheating. arXiv:1912.12294, 2019.

[8] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. arXiv:2205.15997, 2022.

[9] Lu Yue, Dongliang Zhou, Liang Xie, Feitian Zhang, Ye Yan, and Erwei Yin. Safe-VLN: Collision Avoidance for Vision-and-Language Navigation of Autonomous Robots Operating in Continuous Environments. *IEEE Robotics and Automation Letters*, 9(6):4918–4925, 2024.

[10] Junqiao Wang, Zhongliang Yu, Dong Zhou, Jiaqi Shi, and Runran Deng. Vision-Based Deep Reinforcement Learning of UAV Autonomous Navigation Using Privileged Information. arXiv:2412.06313, 2024.

[11] Ran Wei, Joseph Lee, Shohei Wakayama, Alexander Tschantz, Conor Heins, Christopher Buckley, John Carenbauer, Hari Thiruvengada, Mahault Albarracin, Miguel de Prado, Petter Horling, Peter Winzell, and Renjith Rajagopal. Navigation under uncertainty: Trajectory prediction and occlusion reasoning with switching dynamical systems. arXiv:2410.10653, 2024.

[12] Jiuling Zhang. Survey on Monocular Metric Depth Estimation. arXiv:2501.11841, 2025.

[13] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised Learning of Depth and Ego-Motion from Video. In CVPR, pages 6612–6619. IEEE, 2017.

[14] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. arXiv:1907.01341, 2020.

[15] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal Monocular Metric Depth Estimation. arXiv:2403.18913, 2024.

[16] Stefano Gasperini, Nils Morbitzer, HyunJun Jung, Nassir Navab, and Federico Tombari. Robust Monocular Depth Estimation under Challenging Conditions. In ICCV. IEEE, 2023.

[17] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. arXiv:2406.09414, 2024.

[18] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S. Refaat, and Benjamin Sapp. Wayformer: Motion Forecasting via Simple & Efficient Attention Networks. arXiv:2207.05844, 2022.

[19] Mohammad Nazeri, Junzhe Wang, Amirreza Payandeh, and Xuesu Xiao. VANP: Learning Where to See for Navigation with Self-Supervised Vision-Action Pre-Training. In IROS, pages 2741–2746. IEEE, 2024.

[20] Xinhao Liu, Jintong Li, Yicheng Jiang, Niranjan Sujay, Zhicheng Yang, Juexiao Zhang, John Abanes, Jing Zhang, and Chen Feng. CityWalker: Learning Embodied Urban Navigation from Web-Scale Videos. arXiv:2411.17820, 2024.

[21] Steven LaValle. *Planning Algorithms*. Cambridge University Press, 2006.

[22] Xiaojing Zhang, Alexander Liniger, and Francesco Borrelli. Optimization-Based Collision Avoidance. arXiv:1711.03449, 2018.

[23] Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control Barrier Functions: Theory and Applications. arXiv:1903.11199, 2019.

[24] Li Wang, Aaron D. Ames, and Magnus Egerstedt. Safety Barrier Certificates for Collisions-Free Multirobot Systems. *IEEE Transactions on Robotics*, 33(3):661–674, 2017.