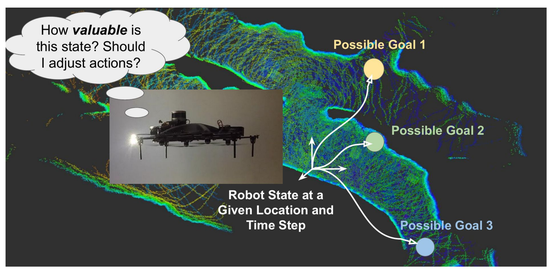
Off-Policy Evaluation with Online Adaptation for Robot Exploration in Challenging Environments

Yafei Hu[1], Junyi Geng[1,2], Chen Wang[1,3], John Keller[1], and Sebastian Scherer[1]
[1] CMU, [2] Penn State, [3] SUNY Buffalo

https://jeffreyyh.github.io/opere/

Background and Motivation

Motivation

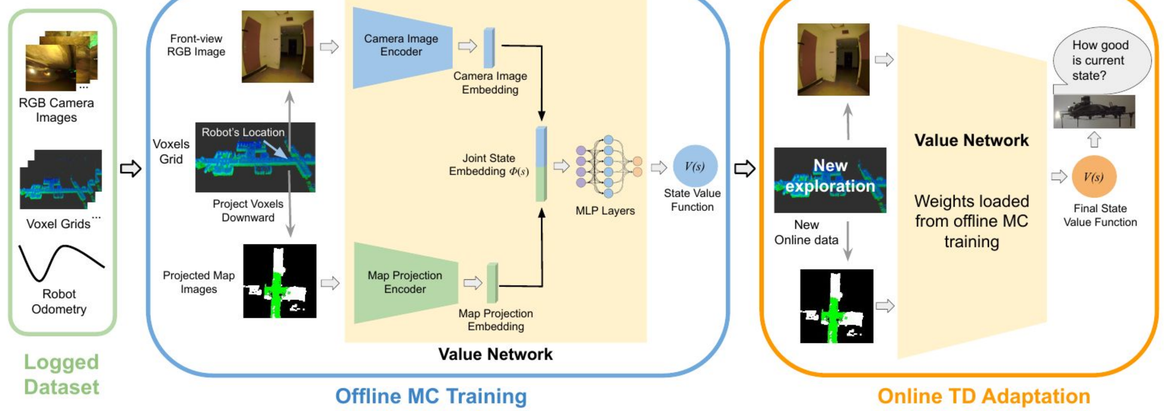


Problem formulation

- Robot exploration as a POMDP
- Learn the value function of states
  - States visited by the robot in the trajectories
  - The goal states that haven't been visited
- Value function learning as an off-policy evaluation (OPE) problem
  - On-policy evaluation for real robots is **costly** and **dangerous**
  - Data collection policy **different** from behaviour policy

Method

Method overview



Reward Relabelling

- Extrinsic reward (sparse)
  - Use number of objects detected at a period of time, devoted as object gain (OG) $OG(t) = O(t) - O(t - \Delta t)$
- Intrinsic rewards:
  - Camera visual coverage map point gain (CG) $CG(t) = C(t) - C(t - \Delta t)$
  - Lidar frontier map point gain (LG) $LG(t) = L(t) - L(t - \Delta t)$
- Final reward $R(t) = aCG(t) + bLG(t) + cOG(t)$

Training and Online Adaptation

Monte-Carlo (MC) Offline Training:

$$J(\boldsymbol{\theta}) = \sum_{s \in \mathcal{S}} \left[V_\pi(s) - \hat{V}(s, \boldsymbol{\theta}) \right]^2$$

$$\boldsymbol{\theta} = \boldsymbol{\theta} + \eta \left[G_t - \hat{V}(\phi(s), \boldsymbol{\theta}) \right] \nabla_{\boldsymbol{\theta}} \hat{V}(\phi(s), \boldsymbol{\theta})$$

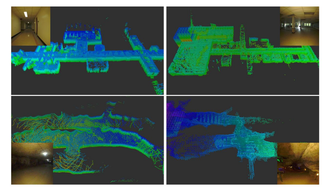
Temporal-Difference (TD) Online Adaptation:

$$\begin{aligned} \boldsymbol{\theta} &= \boldsymbol{\theta} + \eta \left[R(t) + \gamma \hat{V}(\phi(s_{t+1}), \boldsymbol{\theta}) - \hat{V}(\phi(s_t), \boldsymbol{\theta}) \right] \nabla_{\boldsymbol{\theta}} \hat{V}(\phi(s_t), \boldsymbol{\theta}) \\ &= \boldsymbol{\theta} + \eta \left[\mathcal{B} \hat{V}(\phi(s_t), \boldsymbol{\theta}) - \hat{V}(\phi(s_t), \boldsymbol{\theta}) \right] \nabla_{\boldsymbol{\theta}} \hat{V}(\phi(s_t), \boldsymbol{\theta}) \end{aligned}$$

Experiments

Robot Platform

Data Collection Envs

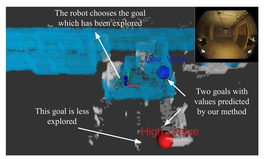


Value Function Estimation Metrics

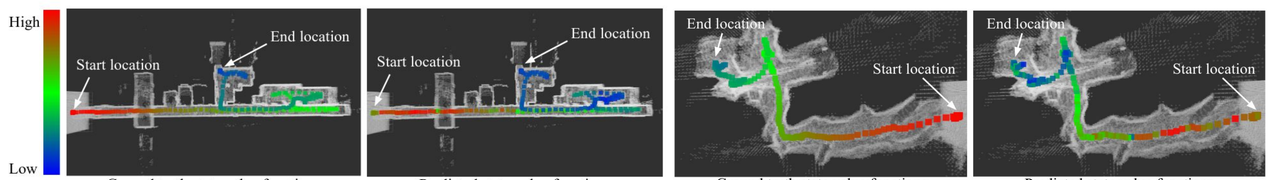
$$\text{NRMSE} = \frac{\text{RMSE}}{\hat{V}(s_t)_{max} - \hat{V}(s_t)_{min}}, \forall t \in [0, T-1]$$

$$R^2 = 1 - \frac{\sum_{t=0}^{T-1} \left[\hat{V}(s_t) - V_\pi(s_t) \right]^2}{\sum_{t=0}^{T-1} \left[\hat{V}(s_t) - \bar{\hat{V}}(s_t) \right]^2}, \bar{\hat{V}}(s_t) = \frac{1}{T} \sum_{t=0}^{T-1} \hat{V}(s_t)$$

$$\text{Regret} = \sum_{t=0}^{T-1} \left[R_C(\pi^*(a_t|o_t)) - R_C(\pi(a_t|o_t)) \right]$$



Qualitative Results of Value Estimation



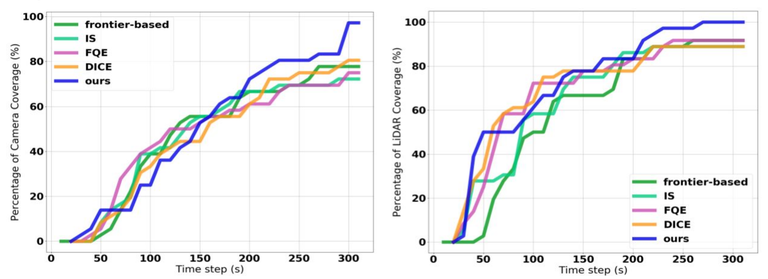
Quantitative Results of Value Estimation

Metrics	Normalized RMSE (↓ the lower the better)				R2 Score (↑ the higher the better)			
Methods	IS [12]	FQE [12]	DICE [32]	Ours	IS [12]	FQE [12]	DICE [32]	Ours
Corridor Env.	0.222±0.000	0.192±0.000	0.176±0.001	**0.129±0.004**	0.563±0.001	0.672±0.001	0.724±0.002	**0.853±0.010**
Room Env.	0.398±0.001	0.506±0.000	0.512±0.000	**0.213±0.002**	-0.593±0.007	-1.573±0.000	-1.634±0.000	**0.543±0.008**
Mine Env.	0.272±0.000	0.264±0.000	0.282±0.001	**0.207±0.002**	0.064±0.003	0.122±0.002	-0.002±0.005	**0.460±0.012**
Cave Env.	0.535±0.000	0.532±0.000	0.535±0.000	**0.164±0.002**	-2.234±0.000	-2.198±0.000	-2.235±0.000	**0.695±0.006**

Regret analysis

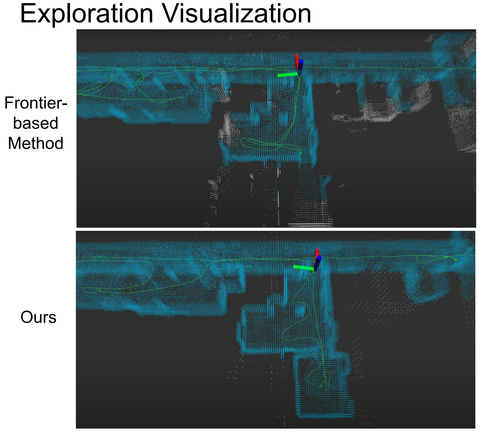
Methods	Corridor Env.	Room Env.	Mine Env.	Cave Env.
Frontier [30]	0.333	0.714	0.625	0.250
IS [12]	0.633	0.443	0.363	0.625
FQE [12]	0.367	0.286	0.263	0.263
DICE [32]	0.617	0.343	0.125	0.375
Ours	**0.108±0.038**	**0.114±0.057**	**0.100±0.050**	**0.013±0.038**

Exploration Performance



Methods	Frontier-based [30]	IS [12]	FQE [12]	DICE [32]	Ours
Camera coverage	77.78%	72.22%	75.00%	80.56%	**97.22%**
LiDAR coverage	91.67%	88.89%	91.67%	88.89%	**100%**

Exploration Visualization



Conclusion

Major takeaways: • We propose an offline MC pre-training and TD online adaptation method to learn value function for robot exploration • The proposed method outperforms other OPE baselines • Real robot testings show certain advantages over frontier-based baseline