

420 **6 Data Collection**

421 In all experiments, we use a legged Boston Dynamics Spot robot and collect robot experi-
 422 ences on eight different types of terrain around the university campus that we labeled as mulch,
 423 pebble sidewalk, cement sidewalk, grass, bushes, marbled rock, yellow bricks, and
 424 red bricks. The data is collected through human teleoperation (by the first and second authors)
 425 such that each trajectory contains a unique terrain throughout, with random trajectory shapes. Note
 426 that STERLING does not require a human expert to teleoperate the robot to collect robot experience
 427 nor does it require the experience to be gathered on a unique terrain per trajectory. We follow this
 428 data collection approach since it is easier to label the terrain for evaluation purposes. STERLING
 429 can also work with random trajectory lengths, with multiple terrains encountered along the same
 430 trajectory, without any semantic labels such as terrain names, and any navigation policy can be used
 431 for data collection. We record 8 trajectories per terrain, each five minutes long, and use 4 trajectories
 432 for training and the remaining for validation.

433 **7 Planning at Deployment**

434 Fig. 5 provides an overview of the cost inference
 435 process for local planning at deployment.
 436 To evaluate the terrain cost $\mathcal{J}_{terrain}(\Gamma)$ for the
 437 constant-curvature arcs, we overlay the arcs on
 438 the bird’s eye view image, extract terrain patches
 439 at states along the arc, and compute the cost ac-
 440 cording to Eq. 2. We compute the visual rep-
 441 resentation, utility value, and terrain cost of all im-
 442 ages at once as a single batch inference. Since the
 443 visual encoder and the utility function are rela-
 444 tively lightweight neural networks with about 0.5
 445 million parameters, we are able to achieve real-
 446 time planning rates of 40 Hz using a laptop-grade
 447 Nvidia GPU.

448 **8 Additional Experiments**

449 In this section, we detail additional experiments
 450 performed to evaluate STERLING-features against
 451 baseline approaches.

452 **8.1 Preference Alignment Evaluation**

453 In addition to the evaluations of STERLING-
 454 features with baseline approaches in five environ-
 455 ments as shown in Sec. 4, we utilize Env. 6 to
 456 further study adherence to operator preferences.
 457 We hypothesize that the discriminative features
 458 learned using STERLING is sufficient to learn the
 459 preference cost for local planning. To test this
 460 hypothesis, in Env. 6 containing three terrains as
 461 shown in Fig. 6, the operator provides two differ-
 462 ent preferences 6(a) and 6(b). While bush is the
 463 least preferred in both cases, in 6(a), sidewalk
 464 is more preferred than grass and in 6(b), both
 465 grass and sidewalk are equally preferred. We see
 466 in Fig. 6 that using STERLING features, the
 planner is able to sufficiently distinguish the
 terrains and reach the goal while adhering to
 operator preferences. Although SE-R [5] adheres
 to operator preference in 6(b), it incorrectly
 maps grass to bush, assigning a higher cost and
 taking a longer

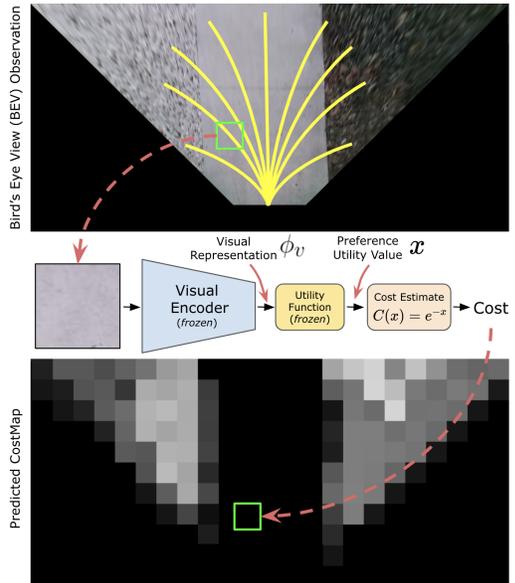


Figure 5: An overview of the cost inference process for local planning at deployment. The constant-curvature arcs (yellow) are overlaid on the BEV image, and the terrain cost $\mathcal{J}_{terrain}(\Gamma)$ is computed on patches extracted along all arcs. White is high cost and black is low cost.



Figure 6: Trajectories traced by different approaches for the task of preference-aligned off-road navigation. Shown here are two different preferences expressed by the operator in the same environment—in 6 (a), sidewalk is more preferred than grass which is more preferred than bush, and in 6 (b), grass and sidewalk are equally preferred and bush is least preferred. We see that without retraining the terrain features, in both cases (a) and (b), STERLING optimally navigates to the goal while adhering to operator preferences.

467 route to reach the goal. On the other hand, RCA [16] fails to adhere to operator preferences since it
 468 directly assigns traversability costs using inertial features.

469 8.2 Evaluating Self-Supervision Objectives

470 In this subsection, we investigate the effective-
 471 ness of STERLING at learning discrimina-
 472 tive terrain features and compare with base-
 473 line unsupervised terrain representation learn-
 474 ing methods such as Regularized Auto-Encoder
 475 (RAE) and SE-R [5]. STERLING uses multi-
 476 modal correlation (\mathcal{L}_{MM}) and viewpoint in-
 477 variance (\mathcal{L}_{VI}) objectives for self-supervised
 478 representation learning, whereas, SE-R and RAE
 479 use soft-triplet-contrastive loss and pixel-wise
 480 reconstruction loss, respectively. Additionally,
 481 we also perform an ablation study on the two
 482 objectives in STERLING to understand their
 483 contributions to learning discriminative terrain
 484 features. To evaluate different visual represen-
 485 tations, we perform unsupervised classification
 486 using k-means clustering and compare their re-
 487 lative classification accuracies with manually la-
 488 beled terrain labels. For this experiment, we
 489 train STERLING, SE-R, and RAE on our training
 490 set and evaluate on a held-out validation set. Fig.
 491 7 shows the results of this study. We see that
 492 STERLING-features using both the self-supervision
 493 objectives perform the best among all methods.
 494 Additionally, we see that using a non-contrastive
 495 representation learning approach such as VICReg [26]
 within STERLING performs better than contrastive
 learning methods such as SE-R, and reconstruction-
 based methods such as RAE. This study shows
 that the proposed self-supervision objectives in
 STERLING indeed help learn discriminative terrain
 features.

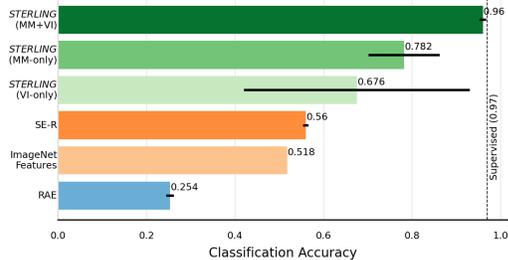


Figure 7: Ablation study depicting classification accuracy (value closer to 1.0 is better) from terrain representations learned using different approaches and objectives. The combined objective (VI + MM) proposed in STERLING achieves the highest accuracy, indicating that the learned representations are sufficiently discriminative of terrains.