

V-STRONG: Visual Self-Supervised Traversability Learning for Off-road Navigation

Sanghun Jung, JoonHo Lee, Xiangyun Meng, Byron Boots, and Alexander Lambert
University of Washington

Abstract—Reliable estimation of terrain traversability is critical for the successful deployment of autonomous systems in wild, outdoor environments. Given the lack of large-scale annotated datasets for off-road navigation, strictly-supervised learning approaches remain limited in their generalization ability. To this end, we introduce a novel, image-based self-supervised learning method for traversability prediction, leveraging a state-of-the-art vision foundation model for improved out-of-distribution performance. Our method employs contrastive representation learning using both human driving data and instance-based segmentation masks during training. We show that this simple, yet effective, technique drastically outperforms recent methods in predicting traversability for both on- and off-trail driving scenarios. We compare our method with recent baselines on both a common benchmark as well as our own datasets, covering a diverse range of outdoor environments and varied terrain types. We also demonstrate the compatibility of resulting costmap predictions with a model-predictive controller. Finally, we evaluate our approach on zero- and few-shot tasks, demonstrating unprecedented performance for generalization to new environments. Videos and additional material can be found here: <https://sites.google.com/view/visual-traversability-learning>.

I. INTRODUCTION

Autonomous navigation in off-road environments requires an accurate understanding of the terrain, particularly in identifying traversable areas by the vehicle. However, contrary to on-road scenarios [1–4], the notion of traversability in unstructured outdoor settings is much more ambiguous, *a priori*. For instance, vehicle interaction with terrain features such as ground-vegetation, rocks, and debris is strongly correlated to their size, shape, and material appearance. As such, manual assignment of traversability labels to off-road perception data is non-trivial and prone to error [5]. Furthermore, the complexity of terrain characteristics and the sheer variety of environments make comprehensive data collection of all terrain types a daunting task. Although efforts have been made towards generating annotated, off-road datasets [6–10], these are generally restricted to a particular set of geospatial locations and seasonal conditions.

For these reasons, achieving reliable performance in traversability prediction via supervised learning remains a challenging problem. Despite the necessity for labeled datasets, recent work on leveraging semantic segmentation has demonstrated the viability of supervised methods for terrain classification [7, 10, 11], typically by generating a Bird’s Eye View (BEV) semantic map of the scene with



Fig. 1: (Top) The Polaris RZR off-road vehicle used for our data collection. (Bottom-left) The vehicle is equipped with multiple RGB-D cameras. (Bottom-middle) Segmentation masks from SAM can disambiguate terrain features. (Bottom-right) Self-supervised learning with mask proposals can achieve robust, fine-grained traversability prediction.

projection of semantic labels [7] or by directly predicting in BEV [10, 11]. Still, the resulting class predictions in such cases must be mapped to a traversability metric or cost, which may require manual tuning and/or additional geometric feature information.

In order to address the challenges of off-road traversability prediction, *self-supervised* learning has shown to be a promising alternative [5, 12–21]. Instead of relying on manually annotated data, many of these approaches use projected traces of vehicle or robot trajectories derived from human-piloted examples as positive traversability labels. Although practical to generate, such datasets are only equipped with positive labels, where much of the observed terrain remains unlabeled. As a result, self-supervised methods that train single-class predictors can be prone to overfitting [5]. Additionally, most of these methods were tested on a restricted class of environments, many of which predominantly consist of on-trail scenarios, which is categorically similar to on-road evaluations [5, 13].

While various sensors (*e.g.*, LiDAR, IMU, etc.) can be used to capture terrain features, RGB cameras offer dense, high-resolution semantic and geometric information. In addition, learning traversability using RGB cameras offers a unique advantage over other sensor modalities: it allows us to leverage large general-purpose models, trained on massive datasets, to derive robust and expressive features from RGB input. Such “Vision Foundation Models” are typically trained on other vision-based tasks, such as semantic segmentation or image classification, but are able to learn rich, mid-level representations that can generalize to other, novel tasks with

impressive results [22–24]. In particular, the Segment Anything Model (SAM) [22], has demonstrated unparalleled zero-shot performance on instance-based semantic segmentation tasks. Trained on a dataset of 11 million images (with over 1.1 billion mask instances), the model is comprised of a Vision-Transformer (ViT) backbone [25], allowing it to retain fine-grained visual information and generate multiple mask proposals for most objects in the scene (Fig. 1). Such models have the potential to drastically improve the generalization performance of traversability learning for off-road autonomy. To date, however, this has yet to be demonstrated. How to best utilize these models remains an open question.

In this paper, we posit that leveraging mask segments for self-supervision offers a simple yet pragmatic method for bootstrapping traversability learning: assuming that image pixels corresponding to the same object or terrain patch should have a similar level of traversability, class-agnostic semantic masks can provide strong priors for self-supervised learning. Specifically, we propose a novel method for pixel-level, contrastive self-supervised learning using projected vehicle trajectories in pixel space along with mask proposals from SAM, overcoming the limitation of solely using trajectory masks as shown in Fig. 2.

We demonstrate the effectiveness of our method on newly collected off-road datasets as our benchmark, on which our method drastically outperforms state-of-the-art baseline methods. Furthermore, while many existing methods validate their approach on off-road but on-trail sequences, ours effectively predicts traversability for both on-/off-trail cases in a number of varied, diverse environments. Lastly, we show how our method can be used for zero- and few-shot traversability learning in new environments not covered in the training data.

II. RELATED WORK

A. Self-supervised Traversability Learning

Given the aforementioned limitations in labeling traversability for off-road terrains, numerous approaches to learning traversability in a self-supervising fashion have been proposed for off-road autonomy [5, 12, 13, 16, 17, 20, 26]. Such efforts have made use of a variety of sensory modalities to provide a self-supervised learning signal. For instance, Castro *et al.* [13] used IMU z -axis measurements as a traversability score, whereas Seo *et al.* [27] used a combination of proprioceptive sensors and LiDAR to extract vehicle trajectory traces.

Recently, several approaches [5, 12, 16, 17, 26] have proposed to predict traversability from visual information. Schmid *et al.* [16] proposed to use autoencoder-based anomaly detection to classify visual terrain features that the system has not traversed. Inspired by [19], they aggregate the vehicle footprints and project them into image space to crop out traversed regions. Then, the autoencoder [28] is trained to reconstruct the traversed regions only. This forces the model to fail on not-traversed areas since they are not observed during the training (*i.e.*, out-of-distribution areas). Thus, at test time, they translate the reconstruction error into the traversability score. However, as the authors remarked, this approach can be

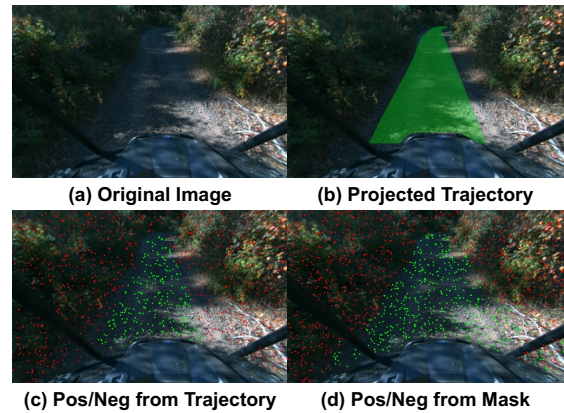


Fig. 2: Illustration of our occlusion handling and trajectory-/mask-based sampling examples. From a given image and trajectory, we use stereo depth to filter out occluded poses and project them into the image space (b). Afterward, we generate random positive samples (bright-green pixels in (c)) and negative samples (red pixels in (c)) within and outside the projected trajectory. Using positive samples from trajectory as query points, we obtain a mask prediction from SAM to cover the whole traversable region. Then, we randomly sample positive and negative points using the mask (d).

susceptible to illumination changes and may produce visual artifacts due to the nature of the reconstruction loss.

B. Contrastive Traversability Learning

Several approaches have adopted contrastive learning in their traversability learning [5, 15, 19, 26]. In Xue *et al.* [15], prototype vectors are learned from embedded positive and unlabeled terrain patches based on LiDAR features, which also serve to generate pseudo-labels for an additional supervised classification task. Additionally, other approaches propose to train a network to generate discriminative feature embeddings either with acoustic features [26] or proprioceptive sensor readings [19] coupled with weakly supervised labels learning to estimate traversability with relatively less manual labeling. In comparison, our method only utilizes visual input from RGB cameras and does not require other sensor measurements.

Perhaps closest in spirit to our work is Seo *et al.* [5], where the authors propose to use Positive-Unlabeled (PU) learning [29] with image-level contrastive learning [30]. The authors adopt a normalizing-flow [31] model and apply a PU learning algorithm for binary classification. Additionally, the contrastive loss [30] is applied to augmented images to encourage the model to have good image representations. However, such a contrastive loss may not be sufficient to provide meaningful information in distinguishing traversable and non-traversable areas. On the contrary, we utilize contrastive learning to promote a model to separate traversable and non-traversable features in the representation space by sampling positive and negative points within and outside the vehicle trajectories.

C. Vision Foundation Models

Recent vision foundation models [22–24] have shown remarkable performance in both accuracy and generalization. In particular, SAM [22] has demonstrated impressive performance in identifying different object instances in high-resolution images, even in unseen environments. As many approaches [32–34] have been proposed using pre-trained

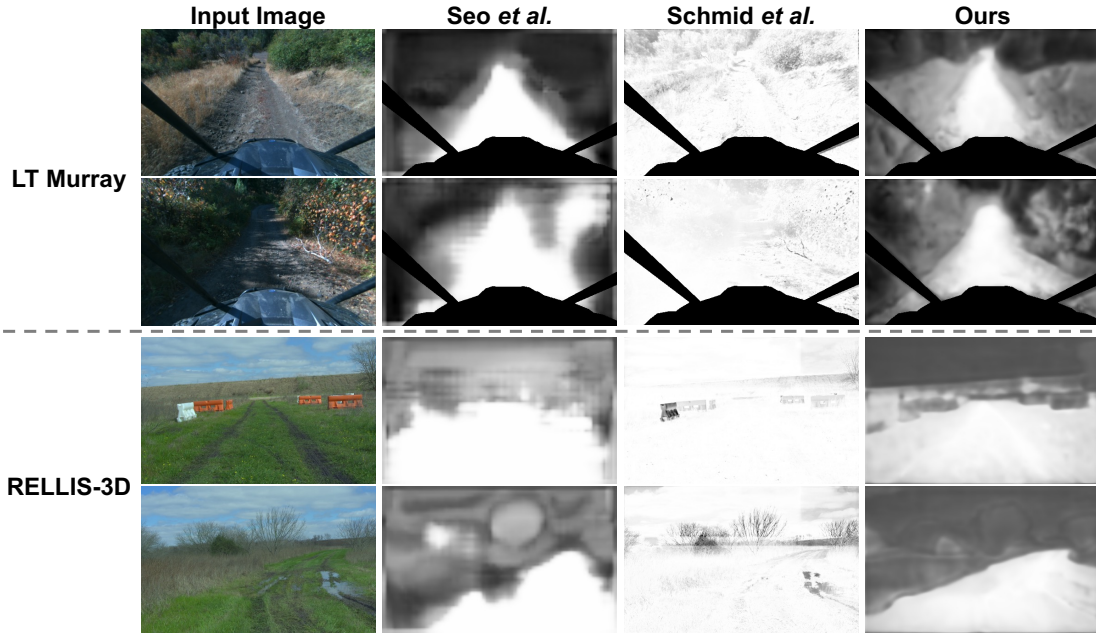


Fig. 4: Qualitative results of Seo *et al.* [5], Schmid *et al.* [16], and our method on RELLIS-3D and LT Murray datasets. We strongly encourage readers to view the supplementary video for more detailed qualitative results.

with pixel-level features $\mathbf{F}_s \in \mathbb{R}^{D \times 1}$ at pixel s . Here, $N = |\mathbb{P}|$, and τ denotes a temperature scalar. We normalize the traversability features \mathbf{F} along with the D dimension before applying the loss.

Finally, our final loss is represented as

$$\mathcal{L} = (1 - \omega_{\text{mask}}) \mathcal{L}_{\text{contra}}(\mathbf{F}, \mathbb{P}_{\text{traj}}, \mathbb{N}_{\text{traj}}) + \omega_{\text{mask}} \mathcal{L}_{\text{contra}}(\mathbf{F}, \mathbb{P}_{\text{mask}}, \mathbb{N}_{\text{mask}}), \quad (4)$$

with weight $\omega_{\text{mask}} \in [0, 1]$.

C. Computing Traversability Cost

While training the model with contrastive losses, we estimate a traversability vector $\mathbf{z} \in \mathbb{R}^D$, for converting traversability features to cost at inference time. Specifically, we update the traversability vector \mathbf{z} using exponential moving averaging (EMA),

$$\mathbf{z} = \alpha \mathbf{z} + (1 - \alpha) \frac{1}{|\mathbb{P}_{\text{traj}}|} \sum_{s \in \mathbb{P}_{\text{traj}}} \mathbf{F}_s, \quad (5)$$

where $\alpha \in [0, 1]$ denotes a momentum value. The vector \mathbf{z} and feature \mathbf{F}_s are normalized before and after the EMA operation. Note that we maintain one traversability vector. This is because the contrastive losses will optimize the traversability decoder to map the SAM feature vectors of traversable regions to have the same representation while distinguishing them from negative feature vectors.

At test time, we convert the traversability features \mathbf{F} to costs $\mathbf{C} \in \mathbb{R}^{H \times W}$ by calculating cosine similarity between \mathbf{F} and \mathbf{z} (*i.e.*, $\mathbf{C} = \mathbf{Fz}$). Since cosine similarity has a range of $[-1, 1]$, the predicted cost \mathbf{C} is always bounded.

IV. EXPERIMENTS

A. Dataset Collection and Processing

Our datasets cover both on-trail and off-trail environments with varied types of terrain. Data collection was performed

at the following sites in the US: LT Murray (WA), Mojave Desert (CA), and California Hills (CA).

LT Murray (WA) sequence is composed of on-trail scenes, where the trails go through dense and sparse vegetation with bushes and trees. We recorded 40 miles of on-trail data.

Mojave Desert (CA) consists of both on-trail and off-trail scenes with rocks, small bushes, Joshua trees, and cacti. We recorded two different runs by following the trail or driving through small bushes and rocks.

CA Hills is composed of off-trail scenes on grassy hills populated with trees. We recorded two different runs by driving the terrain.

We collected images, stereo depths, LiDAR point clouds, and vehicle poses from each run and ran an offline SLAM algorithm [35] to obtain precise state estimations. For evaluation purposes, we labeled LiDAR point clouds and generated segmentation ground truth in BEV. We also set a manual cost per class in BEV for hyper-parameter selection by running real-world experiments.

B. Baselines

For fair evaluations, we compare our method with Seo *et al.* [5] and Schmid *et al.* [16]. Both methods use vehicle trajectories to self-supervise the model to learn traversability. Seo *et al.* adopt a normalizing-flow model [31] along with an image-level contrastive loss to train traversability, while Schmid *et al.* crop the trajectory regions in images and train variational auto-encoder (VAE) [28] to reconstruct the traversed regions. Afterward, the reconstruction error is translated into traversability scores at test time.

C. Implementation details

We adopt the encoder of the ViT-H SAM model as our image encoder and freeze it during training. Note that our full process, from image to prediction, runs around 6 Hz on a single NVIDIA A100 GPU. For training, we use a

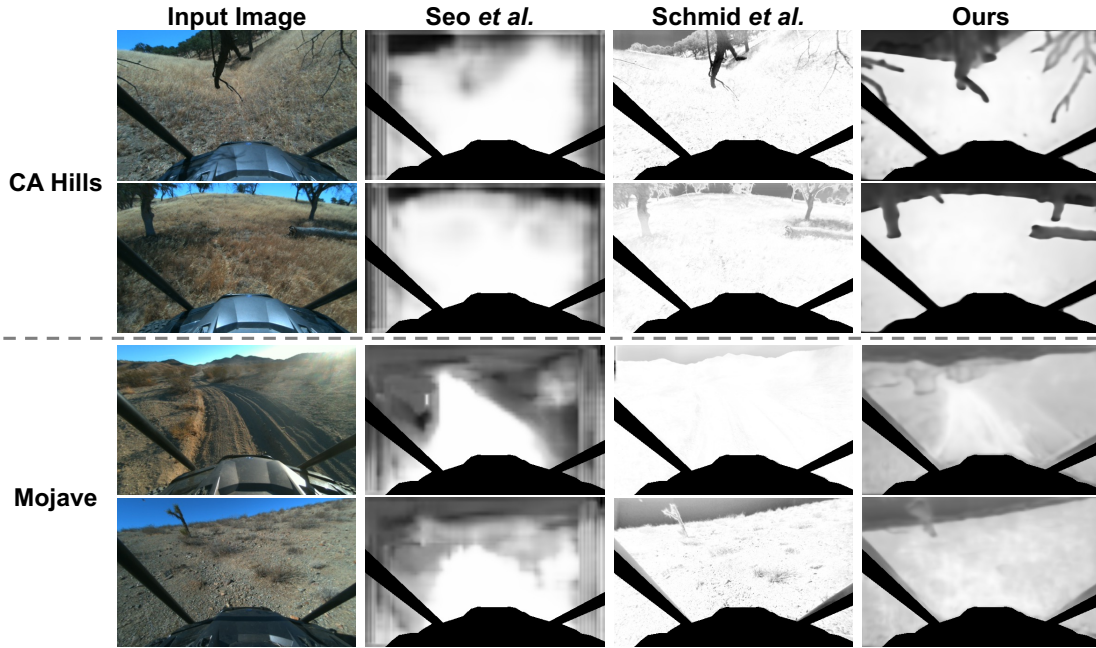


Fig. 5: Qualitative results of Seo *et al.* [5], Schmid *et al.* [16], and our method on CA Hills and Mojave Desert validation sequences. We strongly encourage readers to view the supplementary video for more detailed qualitative results.

learning rate of $1e-3$ and a batch size of 2 and train the model for 1 epoch. We sample 256/1024 positive/negative points for the trajectory contrastive loss, and sample 512/1024 positive/negative points for the mask contrastive loss. Positive/negative samples on the ego vehicle are excluded. We use $\alpha = 0.999$ to update the traversability vector and $T = 300$ for trajectory projection. We set the temperature $\tau = 0.05$ and the mask contrastive loss weight $\omega_{\text{mask}} = 0.05$. These hyper-parameters (*i.e.*, τ and ω_{mask}) have been selected based on the averaged L1 error between predicted costs and manual costs obtained from ground-truth segmentation labels. Note that we project the predicted costs into BEV using stereo depths for hyper-parameter selection since the ground truth labels are in BEV.

D. Qualitative Results

Fig. 4 shows predictions of ours and baseline methods on the RELLIS-3D and LT Murray, which are on-trail datasets. While Seo *et al.* [5] successfully predicts trajectories on trails, it fails to predict lethal objects clearly and distinguish the costs between different semantics such as trees, bushes, and traffic barriers. Schmid *et al.* [16] predicts trails as low costs but fails to mark surrounding lethal objects as high costs. On the other hand, our method clearly marks the trails as low costs and maintains subtle differences well in objects of off-trail terrain.

The strength of our method becomes clearer when it is applied to off-trail scenarios. Fig. 5 illustrates cost prediction results on CA Hills and Mojave Desert sequences, which contain off-trail images. Baselines struggle to find traversable regions correctly or to predict the lethal object as a high cost. Our approach not only distinguishes traversable regions from lethal objects correctly but also assigns different costs for different semantic objects. We conjecture this is because each semantic class has a different frequency to be sampled as

positive. For example, the grass can be sampled as positive if the vehicle goes over it, but the tree cannot be sampled because the vehicle cannot pass through it.

Methods	AUROC \uparrow	AP \uparrow	F1 \uparrow	FPR \downarrow	FNR \downarrow	Pre. \uparrow	Rec. \uparrow
Schmid <i>et al.</i>	0.713	0.666	0.601	0.455	0.273	0.513	0.727
Seo <i>et al.</i>	0.931	0.947	0.888	0.262	0.063	0.844	0.937
Ours	0.959	0.975	0.929	0.100	0.075	0.934	0.925

TABLE I: Evaluation results on the RELLIS-3D dataset [9]. Pre. and Rec. stand for precision and recall.

Methods	Collision Rate \downarrow		
	LT Murray	CA Hills	Mojave Desert
Schmid <i>et al.</i>	0.161	0.106	0.001
Seo <i>et al.</i>	0.128	0.110	0.000
Ours	0.107	0.056	0.001

TABLE II: Collision rates of different methods on the LT Murray, CA Hills, and Mojave Desert datasets.



Fig. 6: Example of the projected costs and MPPI results. The blue trajectory is the expert demonstration, and the red trajectory is a result of successive application of MPPI-optimized control inputs.

E. Quantitative Results

RELLIS-3D dataset We evaluate our method and baselines on the RELLIS-3D dataset [9], which is a publicly available labeled dataset. We adopt AUROC (area under receiver operating characteristic), AUPRC (area under precision-recall curve), F1 score, FPR (false positive rate), FNR (false negative rate), Precision, and Recall metrics for evaluation. Note that we report FPR, FNR, Precision, and Recall metrics at the best threshold that achieves the highest F1 score.

Table I shows the comparisons on the RELLIS-3D dataset. Our method outperforms baselines by a large margin overall, except for FNR and Recall. However, even in such cases,

the gaps are marginal (*i.e.*, 0.012 in both FNR and Recall) compared to our performance gains in other metrics.

Our datasets Since our datasets do not have labels in image space but in BEV, we project the predicted costs into BEV and inpaint the missing costs using the nearest values. Then, we run a model predictive control (MPC) algorithm to obtain the optimal trajectory based on the predicted costmaps (Fig. 6), over all the validation sequences. Specifically, we adopt model predictive path integral (MPPI) [36] as our MPC algorithm and measure the number of collisions with lethal objects over the number of successful MPPI runs (*i.e.*, Collision Rate). To check collision, we use the ground truth labels in BEV and follow the optimal trajectory obtained from running MPPI on predicted costmaps. To achieve a low collision rate, it is important to find all the lethal objects appearing in the image while finding traversable areas correctly. As reported in Table II, our method outperforms other baselines in LT Murray and CA Hills datasets by a large margin. In the case of the Mojave Desert dataset, all three methods show a similar performance achieving nearly 0 collision rates.



Fig. 7: Zero-shot generalization results of our method in LTMurray to Mojave Desert (left) and LTMurray to CA Hills (right) scenarios.



Fig. 8: Few-shot adaptation results of our method in the LTMurray to CA Hills scenario. Iterating two epochs over 10 samples significantly improves the traversability predictions.

Datasets	Collision Rate ↓		
	In-domain	Zero-shot	Few-shot
Mojave Desert	0.001	0.001	-
CA Hills	0.056	0.092	0.048

TABLE III: Evaluation results on the LT Murray, CA Hills, and Mojave Desert datasets.

F. Generalization to New Environments

A strong advantage of having well-trained visual representations is generalization to out-of-distribution environments. To demonstrate the robustness of our method in new environments, we train a model on the on-trail LT Murray dataset and test it on the Mojave Desert and CA Hills. As illustrated in Fig. 7, our method generalizes surprisingly well without any adaptation. The model successfully finds a cactus and a small tree in Mojave and identifies logs lying on the ground in CA Hills, while predicting low cost on traversable areas. The MPPI evaluation results reported in Table III also align with such observations. The few-shot adaptation result in CA Hills outperforms the in-domain results. This demonstrates that our method can effectively transfer from one domain to another, unseen domain.

Methods	AUROC↑	AP↑	F1↑	FPR↓	FNR↓	Pre.↑	Rec.↑
w/o mask	0.920	0.945	0.892	0.240	0.068	0.855	0.932
w/ mask	0.959	0.975	0.929	0.100	0.075	0.934	0.925
Solely w/ mask	0.955	0.972	0.927	0.129	0.063	0.917	0.937

TABLE IV: Effect of applying the mask-based loss to our method in the RELIS-3D dataset. Incorporating mask information significantly improves the overall performance.

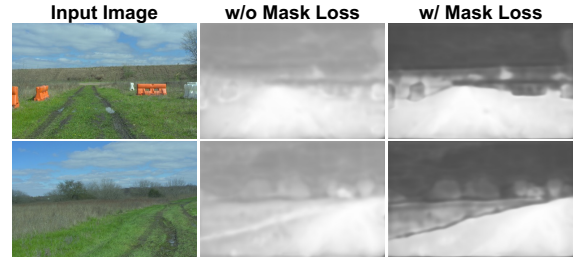


Fig. 9: Qualitative results on the RELIS-3D dataset with and without the mask loss. Incorporating mask information significantly improves discriminability between traversable and non-traversable areas.

G. Ablation Study

Effectiveness of using masks Using vehicle trajectories alone for self-supervision is insufficient since the trajectory cannot cover the whole traversable region. This ablation study aims to reveal how mask-based self-supervision is effective in addressing such a concern by comparing results with and without mask-based loss. As reported in Table IV, applying the mask-based loss brings significant improvements in overall metrics. Additionally, as shown in Fig. 9, the model without mask information overfits vehicle trajectories and marks left and right side regions as high costs even though they are indeed traversable. On the other hand, once we apply the mask-based loss, the model correctly identifies traversable regions and marks them as low costs, mitigating the overfitting problem. Solely using mask-based loss also presents improvements over the one that did not use any mask at all, but it is still behind our final version in most of the metrics. We also observe that false mask proposals from SAM adversely affect the training even though the number of them is very few.

V. CONCLUSION

In this paper, we proposed a novel self-supervised approach to learning traversability in image space using contrastive learning. We show that with the addition of mask-based regularization, guided by robust segmentation proposals, generalization of traversability predictions can be drastically improved and deliver state-of-the-art performance. This is further emphasized by results in out-of-distribution environments. As for future work, we hope to incorporate temporal sequences of image data and learn-able in-painting to improve the quality of predictions, and investigate extending the approach for online adaptation.

VI. ACKNOWLEDGEMENTS

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

REFERENCES

- [1] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5122–5130.
- [2] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.
- [3] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kotschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4990–4999.
- [4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [5] J. Seo, S. Sim, and I. Shim, "Learning off-road terrain traversability with self-supervisions only," *IEEE Robotics and Automation Letters*, 2023.
- [6] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, "Deep multispectral semantic scene understanding of forested environments using multimodal fusion," in *2016 International Symposium on Experimental Robotics*. Springer, 2017, pp. 465–477.
- [7] D. Maturana, P.-W. Chou, M. Uenoyama, and S. Scherer, "Real-time semantic mapping for autonomous off-road navigation," in *Field and Service Robotics: Results of the 11th International Conference*. Springer, 2018, pp. 335–350.
- [8] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5000–5007.
- [9] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "Rellis-3d dataset: Data, benchmarks and analysis," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 1110–1116.
- [10] X. Meng, N. Hatch, A. Lambert, A. Li, N. Wagener, M. Schmittle, J. Lee, W. Yuan, Z. Chen, S. Deng, *et al.*, "Terrainnet: Visual modeling of complex terrain for high-speed, off-road navigation," in *Robotics: Science and Systems*, 2023.
- [11] A. Shaban, X. Meng, J. Lee, B. Boots, and D. Fox, "Semantic terrain classification for off-road autonomous driving," in *Conference on Robot Learning*. PMLR, 2022, pp. 619–629.
- [12] H. Karnan, E. Yang, D. Farkash, G. Warnell, J. Biswas, and P. Stone, "Self-supervised terrain representation learning from unconstrained robot experience," in *ICRA2023 Workshop on Pretraining for Robotics (PT4R)*, 2023.
- [13] M. G. Castro, S. Triest, W. Wang, J. M. Gregory, F. Sanchez, J. G. Rogers, and S. Scherer, "How does it feel? self-supervised costmap learning for off-road vehicle traversability," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 931–938.
- [14] J. Frey, M. Mattamala, N. Chebrolu, C. Cadena, M. Fallon, and M. Hutter, "Fast traversability estimation for wild visual navigation," *arXiv preprint arXiv:2305.08510*, 2023.
- [15] H. Xue, X. Hu, R. Xie, H. Fu, L. Xiao, Y. Nie, and B. Dai, "Contrastive label disambiguation for self-supervised terrain traversability learning in off-road environments," *arXiv preprint arXiv:2307.02871*, 2023.
- [16] R. Schmid, D. Atha, F. Schöller, S. Dey, S. Fakoorian, K. Otsu, B. Ridge, M. Bjelonic, L. Wellhausen, M. Hutter, *et al.*, "Self-supervised traversability prediction by learning to reconstruct safe terrain," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 12 419–12 425.
- [17] E. Chen, C. Ho, M. Maulimov, C. Wang, and S. Scherer, "Learning-on-the-drive: Self-supervised adaptation of visual offroad traversability models," *arXiv preprint arXiv:2306.15226*, 2023.
- [18] L. Wellhausen, R. Ranftl, and M. Hutter, "Safe robot navigation via multi-modal anomaly detection," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1326–1333, 2020.
- [19] L. Wellhausen, A. Dosovitskiy, R. Ranftl, K. Walas, C. Cadena, and M. Hutter, "Where should i walk? predicting terrain properties from images via self-supervised learning," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1509–1516, 2019.
- [20] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. R. Bradski, "Self-supervised monocular road detection in desert terrain." in *Robotics: science and systems*, vol. 38. Philadelphia, 2006.
- [21] J. Zürn, W. Burgard, and A. Valada, "Self-supervised visual terrain classification from unsupervised acoustic feature learning," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 466–481, 2020.
- [22] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.
- [23] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.
- [24] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.
- [25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weis-

- senborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [26] J. Zurn, W. Burgard, and A. Valada, “Self-supervised visual terrain classification from unsupervised acoustic feature learning,” *IEEE Transactions on Robotics*, vol. 37, pp. 466–481, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:208857799>
- [27] J. Seo, T. Kim, K. Kwak, J. Min, and I. Shim, “Scate: A scalable framework for self-supervised traversability estimation in unstructured environments,” *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 888–895, 2023.
- [28] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [29] J. Bekker and J. Davis, “Learning from positive and unlabeled data: A survey,” *Machine Learning*, vol. 109, pp. 719–760, 2020.
- [30] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [31] J. Yu, Y. Zheng, X. Wang, W. Li, Y. Wu, R. Zhao, and L. Wu, “Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows,” *arXiv preprint arXiv:2111.07677*, 2021.
- [32] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, “Track anything: Segment anything meets videos,” *arXiv preprint arXiv:2304.11968*, 2023.
- [33] F. Rajič, L. Ke, Y.-W. Tai, C.-K. Tang, M. Danelljan, and F. Yu, “Segment anything meets point tracking,” *arXiv preprint arXiv:2307.01197*, 2023.
- [34] T. Yu, R. Feng, R. Feng, J. Liu, X. Jin, W. Zeng, and Z. Chen, “Inpaint anything: Segment anything meets image inpainting,” *arXiv preprint arXiv:2304.06790*, 2023.
- [35] W. Hess, D. Kohler, H. Rapp, and D. Andor, “Real-time loop closure in 2d lidar slam,” in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 1271–1278.
- [36] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, “Aggressive driving with model predictive path integral control,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1433–1440.