Active Learning of Robot Vision Using Adaptive Path Planning

Julius Rückin Federico Magistri Cyrill Stachniss Marija Popović

***Abstract*— Robots need robust and flexible vision systems to perceive and reason about their environments beyond geometry. Most of such systems build upon deep learning approaches. As autonomous robots are commonly deployed in initially unknown environments, pre-training on static datasets cannot always capture the variety of domains and limits the robot's vision performance during missions. Recently, self-supervised as well as fully supervised active learning methods emerged to improve robotic vision. These approaches rely on large in-domain pre-training datasets or require substantial human labelling effort. To address these issues, we present a recent adaptive planning framework for efficient training data collection to substantially reduce human labelling requirements in semantic terrain monitoring missions. To this end, we combine high-quality human labels with automatically generated pseudo labels. Experimental results show that the framework reaches segmentation performance close to fully supervised approaches with drastically reduced human labelling effort while outperforming purely self-supervised approaches. We discuss the advantages and limitations of current methods and outline valuable future research avenues towards more robust and flexible robotic vision systems in unknown environments.**

I. Introduction

Perceiving and understanding complex environments is a crucial prerequisite for autonomous systems [1, 2]. In many applications, such as terrain monitoring [3, 4], search and rescue [5, 6], and precision agriculture [7], autonomous robots need to operate in unknown and unseen environments. This poses a major challenge for classical deep learning-based vision systems, which are trained on static datasets and often do not generalise well to new conditions encountered during real-world deployments.

This work examines the problem of semi-supervised active learning to improve robotic vision within an initially unknown environment while minimising human labelling requirements. We tackle this problem by adaptively re-planning the robot's paths online to collect informative training data to re-train its vision system after a mission. We incorporate two sources of labels for network re-training based on the collected data: (i) a human annotator and (ii) automatically generated pseudo labels based on an environment map incrementally built online during a mission.

Active learning is a common approach for reducing human labelling data requirements in computer vision. In the traditional setting, active learning methods select the most informative images from a large, unlabelled dataset [8–11]. The selection criterion is commonly derived based on uncertainty, e.g. using Monte-Carlo dropout [9] or ensembles [12]. These approaches are typically not applicable for robot deployments in unknown environments since the collected data is not known in advance. Thus, recent works investigate combining active learning with robotic planning to guide a robot towards parts of the environment with more informative training data for semantic segmentation [13–15]. A drawback of such methods is that the collected images need to be densely labelled, which is still time- and labour-intensive.

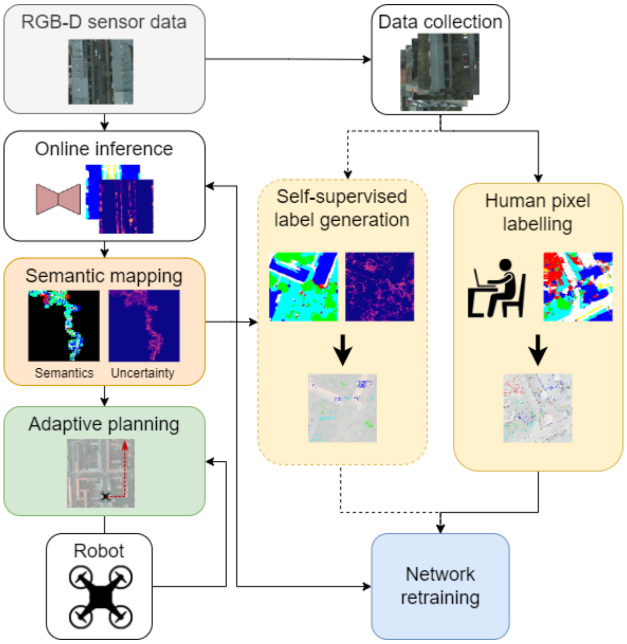

Fig. 1: Our semi-supervised active learning approach in an unknown environment. During a mission, a semantic segmentation network predicts pixel-wise semantics and model uncertainties from an RGB-D image. Both are fused into an uncertainty-aware semantic map, which is used by our adaptive planner to guide the robot towards areas of informative training data where model uncertainty is high. After a mission, the collected data is labelled using two sources of labels: (i) human pixel labelling and (ii) self-supervised pseudo label generation from the semantic map.

Conversely, self-supervised active learning methods automatically generate pseudo labels from maps incrementally built during a mission [16–18], without relying on human labelling. However, their applicability to diverse sets of unknown environments is limited since they require large labelled in-domain pre-training datasets to produce high-quality pseudo labels without systematic prediction errors. These pre-training requirements are typically hard to realise in real-world robotic deployment settings, e.g. outdoor and aerial monitoring, where training data is scarce.

J.R., F.M., and C.S. are with the Institute of Geodesy and Geoinformation, Cluster of Excellence PhenoRob, University of Bonn. M.P. is with MAVLab, Faculty of Aerospace Engineering, TU Delft. C.S. is also with the University of Oxford, UK, and the Lamarr Institute for Machine Learning and Artificial Intelligence, Germany. This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 (PhenoRob). Corresponding: `jrueckin@uni-bonn.de`.

Our paper bridges the gap between these two streams of research. We start with a semi-supervised adaptive path planning framework for robotic active learning, introduced in our recent journal publications [15, 19]. As illustrated in Fig. 1, the approach combines automatically generating uncertainty-aware self-supervised pseudo labels from a semantic map and selecting informative human-labelled training data. We explore sparse human label selection techniques to further reduce labelling requirements [20, 21]. For adaptive planning, our approach maintains an uncertainty-aware semantic map, enabling us to guide the robot to collect images for labelling from high-uncertainty areas. By combining human and pseudo labels, our goal is to maximise semantic segmentation performance while reducing human labelling effort compared to previous fully supervised works in robotic active learning. Based on our findings, our paper concludes with a new and previously unpublished discussion of limitations and open challenges to drive the research community supporting label-efficient robotic learning paradigms.

II. Our Approach

Considering a robot equipped with an RGB-D sensor, we present an approach for collecting images in an unknown environment to improve semantic segmentation with minimal human labelling effort [15, 19] as depicted in Fig. 1.

A. Probabilistic Semantic Environment Mapping

A crucial requirement for pseudo label generation and adaptive planning is a probabilistic map capturing information about the environment. We use probabilistic multi-layered semantic environment mapping to fuse semantic model predictions. The environment is discretised into two voxel maps $\mathcal{M}_S : V \rightarrow \{0,1\}^{K \times W \times L \times H}$ and $\mathcal{M}_U : V \rightarrow [0,1]^{W \times L \times H}$ defined over $W \times L \times H$ spatially independent voxels V. The semantic map $\mathcal{M}_S$ consists of K layers with one layer per class and is recursively updated using occupancy grid mapping [22]. The model uncertainty map $\mathcal{M}_U$ is updated using maximum likelihood estimation. Additionally, we maintain a count map $\mathcal{M}_T : V \rightarrow \mathbb{N}^{W \times L \times H}$ to track the occurrences in the human-labelled training data utilised in our planning objective. The semantic predictions and model uncertainties change as the semantic segmentation model is re-trained after each robot mission. Thus, we re-compute the semantic and model uncertainty maps after model re-training using previously collected RGB-D images to obtain maximally up-to-date map priors for adaptive planning.

B. Adaptive Informative Path Planning

We aim to maximise the performance of a semantic segmentation model with minimal human labelling effort after re-training it on the collected training data. Our map-based global planning methods search for a path $\psi^* = (\mathbf{p}_1, \ldots, \mathbf{p}_N) \in \Psi$ with a variable number $N \in \mathbb{N}$ of robot poses $\mathbf{p}_i \in \mathbb{R}^D$, $i \in \{1, \ldots, N\}$, in the set of potential paths Ψ, that maximises an information criterion $I : \Psi \rightarrow \mathbb{R}_{\geq 0}$:

$$\psi^* = \underset{\psi \in \Psi}{\operatorname{argmax}}\, I(\psi), \text{ s.t. } C(\psi) \leq B \,, \tag{1}$$

where I assigns an information value to each possible path $\psi \in \Psi$, $B \geq 0$ is the mission budget, and $C : \Psi \rightarrow \mathbb{R}_{\geq 0}$ defines the required budget to execute the path ψ.

At each time step t, we adaptively re-plan the path ψ_t^* based on the current map states $\mathcal{M}_U^t$ and $\mathcal{M}_T^t$, and execute the next-best pose $\mathbf{p}_{t+1}$ to collect informative training data. The information criterion estimates the effect of a candidate training image recorded at pose $\mathbf{p}$ on a semantic segmentation model's performance. To this end, our information criterion I trades off between model uncertainty and training data diversity. Based on the camera's field of view, we compute a set of voxels $V_\mathbf{p}$ visible from pose $\mathbf{p}$ and extract currently mapped model uncertainties $\mathcal{M}_U^t(v)$ and training data occurrences $\mathcal{M}_T^t(v)$ for all voxels $v \in V_\mathbf{p}$. Pose $\mathbf{p}$ contains high information value if model uncertainties $\mathcal{M}_U^t(v)$ are high while training data occurrences $\mathcal{M}_T^t(v)$ are low. To foster exploration, voxels v in unknown space receive a constant exploration bonus $\mathcal{M}_U^t(v) = c_u$, where $c_u > 0$.

C. Efficient Labelling

We propose a semi-supervised training strategy for improving the robot's semantic vision. We utilise a semantic segmentation network to predict the pixel-wise probabilistic semantic labels of images. To maximise model performance, we combine human-labelled and automatically pseudo-labelled images during network training.

Combining ideas from Shin et al. [20] and Xie et al. [21], we propose a new model architecture-agnostic pixel selection procedure for sparse human labels that trades off between label informativeness and diversity. After each mission, we predict each pixel's maximum likelihood semantic label and compute its region impurity score following Xie et al. [21]. A pixel's region impurity and, thus, its information value upon re-training is high whenever the number of different classes predicted within its neighbourhood is high, as semantics are usually locally non-cluttered. We select the $\beta\,\%$ pixels with the highest region impurity to ensure an information value lower bound. Then, we sample α pixels uniformly at random from these $\beta\,\%$ pixels to foster training data diversity.

Similarly to self-supervised robotic active learning approaches [16, 17], we use our incrementally online-built semantic and model uncertainty maps (Sec. II-A) to generate pseudo labels. Given a pose, we render pixel-wise maximum likelihood semantic pseudo labels and model uncertainties from these maps. After each mission, for all images collected in any of the previous missions from respective robot poses, we (re-)render pseudo labels and model uncertainties based on the most recent map beliefs. In contrast to previous works [16, 17], we only use a sparse set of α pseudo-labelled pixels per image as we experimentally found that sparse pseudo labels balance the human and self-supervision best. Building upon Shin et al. [20], for each image, we select the $\beta\%$ pixels with the lowest map-based model uncertainties to ensure a lower bound on the pseudo label quality. Then, we sample α pixels uniformly at random from these $\beta\,\%$ pixels to foster diversity of the sparse pseudo labels.

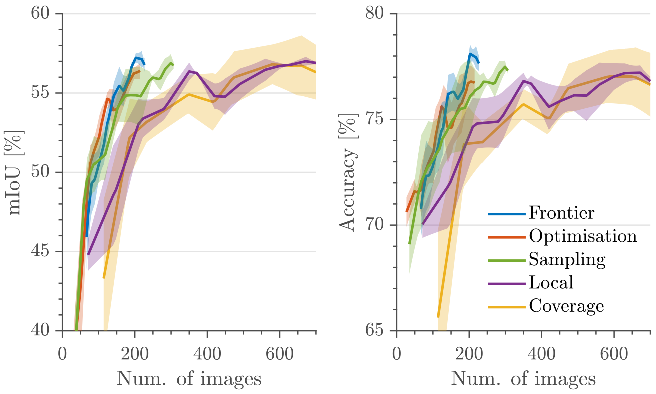

Fig. 2: Our global map-based adaptive planners (blue, orange, green) compared to state-of-the-art local planning (purple) and classical non-adaptive coverage paths (yellow). Our map-based planners require substantially fewer pixel-wise human-labelled images to reach the same performance as coverage and local planning.

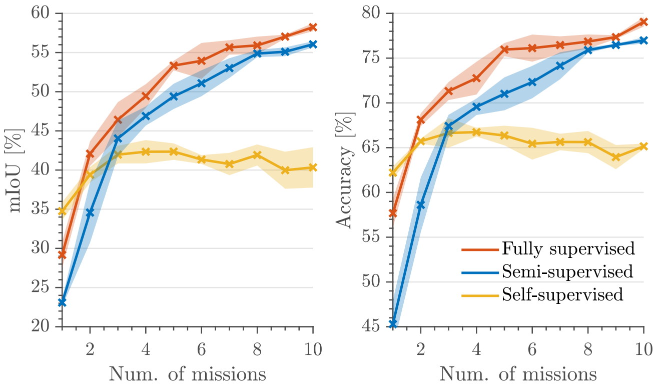

Fig. 3: Our semi-supervised adaptive frontier planning compared to fully and self-supervised adaptive frontier planning. Our semi-supervised approach almost reaches the fully supervised performance while clearly outperforming the self-supervised approach.

III. Experimental Results

We evaluate our framework on the real-world ISPRS Potsdam orthomosaic dataset [23] and simulate 10 UAV missions from 30 m altitude with a mission budget of 1800 s. The UAV uses a downwards-facing RGB-D camera with a footprint of 400 px×400 px. In this work, we consider four adaptive planners [15] to optimise our planning objective proposed in Sec. II-B and a standard coverage pattern:

Local is an image-based planner locally following the direction of the highest training data information in the image recorded at the current UAV position. This planner resembles the state-of-the-art method by Blum et al. [13];

Frontier is a global map-based geometric planner guiding the UAV towards frontiers of explored and unexplored terrain with the highest training data information;

Optimisation selects a path over a fixed horizon of multiple time steps to optimise the path's overall training data information following the work by Popović et al. [7];

Sampling utilises Monte-Carlo tree search (MCTS) [24] to find the next position that maximises the future training data information in a sampling-based fashion.

We use Bayesian ERFNet [14] pre-trained on the Cityscapes dataset [25]. Re-training after each mission starts from this checkpoint and stops after convergence on the validation set. We use a one-cycle learning rate, a batch size of 8, and weight decay $\lambda = (1-p)/2N$, where $p = 0.5$ is the dropout probability and N is the number of images [9].

In Fig. 2, we evaluate the performance of the adaptive planners against a traditional pre-planned coverage-based training data collection. We report the mean Intersection-over-Union (mIoU) and accuracy over the number of pixel-wise human-labelled training images averaged over three different UAV starting locations. Higher semantic segmentation performance, thanks to newly added images, indicates better active learning and, thus, planning performance. The local planner, on average, does not perform better than the coverage baseline. All our adaptive map-based planners, on average, reach higher active learning performance than the coverage baseline (yellow) and local planner (purple) with substantially fewer human-labelled images. Specifically, the frontier planner (blue) requires approx. 200 images to reach the performances of the coverage planner on approx. 600 images. These results verify that our global map-based adaptive planners outperform classical pre-planned data collection campaigns for active learning in semantic terrain mapping missions. Further, they show that our map-based adaptive planners reach higher active learning performance than previous state-of-the-art local planning [13].

In Fig. 3, we select the map-based adaptive frontier planner to evaluate the effect of our proposed semi-supervised training data labelling strategy. We report the mIoU and accuracy after each mission's network re-training averaged over three different runs to account for the inherent randomness in the sparse pixel selection procedure. We compare our semi-supervised labelling strategy (blue) to (i) fully supervised pixel-wise human labels (orange), and (ii) to a purely self-supervised labelling strategy (yellow) pre-trained on a small set of human-labelled Potsdam ISPRS images and rendering pixel-wise pseudo labels based on the current semantic map belief. Our semi-supervised strategy performs almost on par with the fully supervised human labelling while requiring approx. only 0.5% of the human-labelled pixels. Interestingly, the self-supervised approach fails to improve model performance after four re-deployments. This indicates that efficiently selecting human-labelled pixels is a key ingredient of our framework to circumvent reinforced self-supervision errors in semantic terrain mapping missions.

IV. Discussion & Future Directions

Next, we discuss the open challenges in adaptive robotic planning for active learning of robust vision and suggest future research directions to address them.

A. Faster to Answer Human Labelling Queries

Although self-supervised methods do not require human annotations to improve vision performance, these approaches rely on large human-labelled pre-training datasets containing data similar to deployment. Thus, self-supervised methods are often upper-bounded in performance by the pre-training and domain shift during deployment. In contrast, fully supervised methods induce substantial human labelling costs requiring pixel-wise annotations [13–15]. Our results show that, for active learning in semantic terrain mapping, combining sparse human labels and self-supervision enables reducing the number of human-labelled pixels to approx. 0.5% of fully supervised methods while maintaining performance [19]. Although recent studies suggest that sparse pixel selection reduces annotation time [20, 26], human labelling query costs are, from our perspective, still too high to be easily and repeatedly answered by an operator. One idea could be to explore uncertainty-guided one-click annotations [27, 28]. Another promising path could be to leverage foundation models, such as SAM [29], and prompt them in a targeted, potentially uncertainty-aware fashion.

B. Novel Embodied Self-supervised Learning Methods

Human-guided methods still suffer from costly human annotations [13, 15, 19]. High-quality self-supervised labels are required to keep the human labelling effort low and reach maximal prediction performance. To this end, self-supervised methods create pseudo labels from an online-built semantic map [16–19]. These methods render pseudo labels from voxel-based maps at viewpoints encountered during deployment. However, voxel-based maps cannot render image-label pairs from novel viewpoints. Semantic neural rendering approaches recently enhanced self-supervised pseudo labels, rendering high-quality image-label pairs from novel viewpoints, outperforming voxel map-based pseudo label generation [30]. Combining neural rending methods with adaptive planning could improve current systems without additional human labels. Further, robotic active learning methods leverage the robot's embodiment in the environment using adaptive planning to enhance spatial consistency of pseudo labels [17, 18]. Most methods use generated map-based pseudo labels directly with standard loss functions during network training [16, 17], but they do not leverage advanced self-supervised methods, such as contrastive learning. Chen et al. [31] show that additionally enforcing spatial consistency during network training using contrastive learning techniques improves object-goal navigation. These advanced self-supervised techniques could also improve active learning for robotic vision systems.

C. Improved Uncertainty Quantification

As discussed by Chaplot et al. [18], overconfidently wrong predictions reinforce prediction errors after re-training on these predictions in a self-supervised fashion. Even human-guided methods require well-calibrated uncertainty estimation to create informative human labelling queries that maximise performance while minimising labelling effort [15]. Thus, better-calibrated model uncertainty estimation techniques are required as current techniques tend to produce overconfident predictions [12, 32, 33]. Further, current methods ignore various sources of uncertainty. All methods use some measure of model uncertainty or confidence [15, 17–19] to collect potentially informative new training data. Future research could integrate and disentangle other sources of uncertainty, such as data uncertainty [34] induced by environmental factors or noisy sensors. This information could be used to avoid requesting human labels for inputs with high data uncertainty that contribute little to the model improvements [34] or to adaptively plan novel viewpoints that might reduce these uncertainties [35].

D. Towards Continual Active Learning

Another key challenge for efficient learning of robotic vision systems is the robot's ability to continually learn about new unseen environments while transferring the knowledge gained during previous deployments [36] without suffering from catastrophic forgetting [37]. This problem of continual learning is largely ignored in robotic active learning methods. To the best of our knowledge, Frey et al. [16] proposed the only method for continual active learning using experience replay [38]. However, they do not leverage adaptive planning for training data collection. Further, although conceptually simple and effective against catastrophic forgetting, experience replay is storage- and compute-inefficient as its complexity scales linearly with the number of deployments. Combining adaptive replanning with continual learning over sequential deployments in various environments could lead to more robust vision systems and a more targeted continuous collection of informative training data while leveraging already gained previous knowledge.

E. Improved Model Re-training Efficiency

Similarly to continual active learning, current methods for active learning within a single environment require iterative network re-training to adapt training data collection based on previously collected data. Although most methods use lightweight networks for improved training and inference speed [15, 17, 19], iterative re-training is prohibitively expensive in applications that require fast online adaption of vision or re-deployment cycles. One way to improve the network re-training efficiency could be to leverage vision foundation models [29] as pre-trained feature extractors combined with small, trainable adapter networks. This could mitigate the costly re-training of larger networks while allowing the robot vision to profit from few-shot generalisation.

V. Conclusion

We presented our adaptive planning approach for semi-supervised active learning of robotic vision in unknown environments [19]. Our experimental results show that our semi-supervised approach outperforms traditional pre-planned data collection campaigns and purely self-supervised robotic active learning approaches in semantic terrain monitoring missions. Further, our approach requires only approx. 0.5% of the human-labelled pixels of fully supervised robotic active learning methods [15] while maintaining semantic segmentation performance. We conclude with a discussion of open challenges and identify future directions to advance state-of-the-art robotic active learning methods.

References

[1] G. Lenczner, A. Chan-Hon-Tong, B. Le Saux, N. Luminari, and G. Le Besnerais, “DIAL: Deep Interactive and Active Learning for Semantic Segmentation in Remote Sensing,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 3376–3389, 2022.

[2] G. Georgakis, B. Bucher, K. Schmeckpeper, S. Singh, and K. Daniilidis, “Learning to Map for Active Semantic Goal Navigation,” in *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2022.

[3] R. Marchant, F. Ramos, and S. Sanner, “Sequential Bayesian optimisation for spatial-temporal monitoring,” in *Proc. of the Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2014.

[4] G. Hitz, E. Galceran, M.-È. Garneau, F. Pomerleau, and R. Siegwart, “Adaptive Continuous-Space Informative Path Planning for Online Environmental Monitoring,” *Journal of Field Robotics (JFR)*, vol. 34, no. 8, pp. 1427–1449, 2017.

[5] F. Niroui, K. Zhang, Z. Kashino, and G. Nejat, “Deep reinforcement learning robot for search and rescue applications: Exploration in unknown cluttered environments,” *IEEE Robotics and Automation Letters (RA-L)*, pp. 610–617, 2019.

[6] J. L. Baxter, E. Burke, J. M. Garibaldi, and M. Norman, “Multi-robot search and rescue: A potential field based approach,” *Autonomous Robots*, pp. 9–16, 2007.

[7] M. Popović, T. Vidal-Calleja, G. Hitz, J. J. Chung, I. Sa, R. Siegwart, and J. Nieto, “An Informative Path Planning Framework for UAV-based Terrain Monitoring,” *Autonomous Robots*, vol. 44, no. 6, pp. 889–911, 2020.

[8] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, “Selective Sampling Using the Query by Committee Algorithm,” *Machine Learning*, vol. 28, no. 2, pp. 133–168, 1997.

[9] Y. Gal, R. Islam, and Z. Ghahramani, “Deep Bayesian Active Learning with Image Data,” in *Proc. of the Int. Conf. on Machine Learning (ICML)*, 2017.

[10] O. Sener and S. Savarese, “Active Learning for Convolutional Neural Networks: A Core-Set Approach,” in *Proc. of the Int. Conf. on Learning Representations (ICLR)*, 2018.

[11] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, “Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation,” in *Proc. of the Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, 2017.

[12] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, “The Power of Ensembles for Active Learning in Image Classification,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] H. Blum, S. Rohrbach, M. Popović, L. Bartolomei, and R. Siegwart, “Active Learning for UAV-based Semantic Mapping,” in *Proc. of Robotics: Science and Systems Workshop on Informative Path Planning and Adaptive Sampling*, 2019.

[14] J. Rückin, L. Jin, F. Magistri, C. Stachniss, and M. Popović, “Informative Path Planning for Active Learning in Aerial Semantic Mapping,” in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2022.

[15] J. Rückin, F. Magistri, C. Stachniss, and M. Popović, “An Informative Path Planning Framework for Active Learning in UAV-Based Semantic Mapping,” *IEEE Trans. on Robotics (TRO)*, vol. 39, no. 6, pp. 4279–4296, 2023.

[16] J. Frey, H. Blum, F. Milano, R. Siegwart, and C. Cadena, “Continual Adaptation of Semantic Segmentation using Complementary 2D-3D Data Representations,” *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 4, pp. 11 665–11 672, 2022.

[17] R. Zurbrügg, H. Blum, C. Cadena, R. Siegwart, and L. Schmid, “Embodied Active Domain Adaptation for Semantic Segmentation via Informative Path Planning,” *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 4, pp. 8691–8698, 2022.

[18] D. S. Chaplot, M. Dalal, S. Gupta, J. Malik, and R. R. Salakhutdinov, “SEAL: Self-supervised Embodied Active Learning using Exploration and 3D Consistency,” *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2021.

[19] J. Rückin, F. Magistri, C. Stachniss, and M. Popović, “Semi-Supervised Active Learning for Semantic Segmentation in Unknown Environments Using Informative Path Planning,” *IEEE Robotics and Automation Letters (RA-L)*, vol. 9, no. 3, pp. 2662–2669, 2024.

[20] G. Shin, W. Xie, and S. Albanie, “All you need are a few pixels: semantic segmentation with pixelpick,” in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2021.

[21] B. Xie, L. Yuan, S. Li, C. H. Liu, and X. Cheng, “Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[22] H. Moravec and A. Elfes, “High resolution maps from wide angle sonar,” in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 1985.

[23] ISPRS. (2018) 2D Semantic Labeling Contest. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/semantic-labeling.aspx

[24] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, “A Survey of Monte Carlo Tree Search Methods,” *IEEE Trans. on Computational Intelligence and AI in Games*, vol. 4, no. 1, pp. 1–43, 2012.

[25] 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/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] R. Benenson and V. Ferrari, “From colouring-in to pointillism: revisiting semantic segmentation supervision,” *arXiv preprint arXiv:2210.14142*, 2022.

[27] N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang, “Deep interactive object selection,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[28] S. Majumder, A. Khurana, A. Rai, and A. Yao, “Multi-stage fusion for one-click segmentation,” in *Proc. of the German Conf. on Pattern Recognition (GCPR)*, 2020.

[29] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, “Segment Anything,” in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2023.

[30] Z. Liu, F. Milano, J. Frey, R. Siegwart, H. Blum, and C. Cadena, “Unsupervised continual semantic adaptation through neural rendering,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[31] B. Chen, J. Kang, P. Zhong, Y. Liang, Y. Sheng, and J. Wang, “Embodied Contrastive Learning with Geometric Consistency and Behavioral Awareness for Object Navigation,” in *ACM Int. Conf. on Multimedia (ACM MM), (Accept)*, 2024.

[32] J. Postels, M. Segu, T. Sun, L. D. Sieber, L. Van Gool, F. Yu, and F. Tombari, “On the practicality of deterministic epistemic uncertainty,” in *Proc. of the Int. Conf. on Machine Learning (ICML)*, 2022.

[33] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” in *Proc. of the Int. Conf. on Machine Learning (ICML)*, 2016.

[34] A. Kendall and Y. Gal, “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” *Proc. of the Conf. on Neural Information Processing Systems (NIPS)*, 2017.

[35] D. Morilla-Cabello, J. Westheider, M. Popović, and E. Montijano, “Perceptual factors for environmental modeling in robotic active perception,” in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2024.

[36] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, and N. Díaz-Rodríguez, “Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges,” *Information Fusion*, vol. 58, pp. 52–68, 2020.

[37] M. McCloskey and N. J. Cohen, “Catastrophic interference in connectionist networks: The sequential learning problem,” in *Psychology of learning and motivation*. G. H. Bower, Ed. Academic Press, 1989, vol. 24, pp. 109–165.

[38] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, “Experience replay for continual learning,” in *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2019.