# Semi-Supervised Active Learning for Semantic Segmentation in Unknown Environments Using Informative Path Planning

Julius Rückin Federico Magistri Cyrill Stachniss Marija Popović

***Abstract*—Semantic segmentation enables robots 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 perception performance during missions. Recently, self-supervised and 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. We propose a planning method for semi-supervised active learning of semantic segmentation that substantially reduces human labelling requirements compared to fully supervised approaches. We leverage an adaptive map-based planner guided towards the frontiers of unexplored space with high model uncertainty collecting training data for human labelling. A key aspect of our approach is to combine the sparse high-quality human labels with pseudo labels automatically extracted from highly certain environment map areas. Experimental results show that our method reaches segmentation performance close to fully supervised approaches with drastically reduced human labelling effort while outperforming self-supervised approaches.**

***Index Terms*—Motion and Path Planning, Deep Learning for Visual Perception, Semantic Scene Understanding**

## I. Introduction

PERCEIVING and understanding complex environments is a crucial prerequisite for autonomous systems [1, 2]. At the same time, robots are increasingly utilised in diverse terrains to execute various tasks, such as monitoring [3, 4], search and rescue [5, 6], and precision agriculture [7]. Thus, robotic perception systems need to adapt to novel domains and terrains. However, classical deep learning-based semantic segmentation systems are pre-trained on static datasets that often fall short in covering the varying domains and semantics encountered during real-world robot deployments.

This work examines the problem of semi-supervised active learning to improve robotic vision within an initially unknown environment. We aim to maximise the robot's semantic segmentation performance while minimising human labelling requirements. The robot re-plans paths online to collect informative training data to re-train a semantic segmentation model 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 a semantic environment map incrementally built online during a mission.

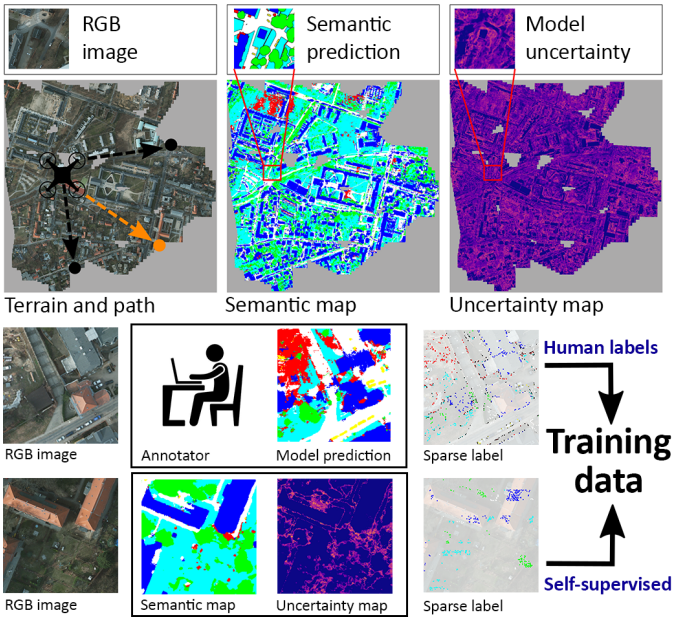


Fig 1: Our semi-supervised active learning approach in an unknown environment (top). We infer semantic segmentation (top-centre) and model uncertainty (top-right) and fuse both in environment maps. The robot re-plans its path (orange, top-left) to collect diverse uncertain images. After each mission, we select sparse sets of pixels for human and self-supervised labelling (bottom). Self-supervised labels are rendered from low-uncertainty semantic map regions. Human labels are queried for regions of cluttered model predictions.

To reduce human labelling effort, active learning methods select the most informative images from a static pool of unlabelled data [8–11]. Recent works combine active learning with robotic planning to reduce the amount of labelled training data in unknown environments [12–14]. However, these methods require time-consuming dense pixel-wise human-labelled images to train semantic segmentation models. In parallel, self-supervised active learning methods automatically generate pseudo labels from semantic maps incrementally built during

Manuscript received: Dec 07, 2023; Accepted: Jan 22, 2024. This paper was recommended for publication by Editor Aniket Bera upon evaluation of the Associate Editor and Reviewers' comments.

All authors are with the University of Bonn, Cluster of Excellence PhenoRob, Institute of Geodesy and Geoinformation. Cyrill Stachniss is also with the University of Oxford and 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.

Digital Object Identifier (DOI): see top of this page.

a mission [15–17]. Although these approaches do not rely on human labels, their applicability to unknown environments is limited since they require large labelled in-domain pre-training datasets to produce high-quality pseudo labels without systematic prediction errors.

The main contribution of this paper is a novel semi-supervised informative path planning approach for robotic active learning. Our approach bridges the gap between the general applicability of fully supervised methods and low human labelling requirements of self-supervised methods. A key novelty of our adaptive planning method is combining the selection of sparse and informative human-labelled training data and automatically generating highly certain pseudo labels as shown in Fig. 1. We fuse semantic predictions and Bayesian model uncertainty estimates [18] into environment maps. Based on the model uncertainty map, our planner adaptively collects images from high-uncertainty areas. Following recent works in semi-supervised learning, we select only a sparse set of to-be-labelled informative pixels from each image [19, 20]. To further improve model performance, we automatically render highly certain pseudo labels based on the semantic and model uncertainty maps. By combining human and pseudo labels, we aim to maximise semantic segmentation performance while reducing human labelling effort compared to previous works in robotic active learning.

In sum, we make the following three key claims. First, our approach drastically reduces the number of human-labelled pixels compared to fully supervised active learning approaches while preserving similar semantic segmentation performance and outperforming self-supervised methods. Second, selecting sparse human labels in a targeted way improves semantic segmentation performance while minimising overall human labelling efforts. Third, the uncertainty-aware generation of pseudo labels further improves semantic segmentation performance compared to using human labels only. We will open-source our code for usage by the community at: https://github.com/dmar-bonn/ipp-ssl.

## II. Related Work

Our approach combines computer vision research aiming to minimise human labelling effort for training semantic segmentation models and informative path planning.

**Active learning** aims to select a minimal subset of informative to-be-labelled training data from a pool of unlabelled data that maximises model performance. Some works estimate uncertainty to access a sample's information value [9, 21] utilising methods such as Monte-Carlo dropout [9] or ensembles [21]. These strategies select samples from a large pool of unlabelled data. In contrast, our planning method exploits Bayesian model uncertainty estimates [9] fused into an environment map guiding the robot towards high-uncertainty areas to incrementally collect samples during a mission.

Shin et al. [20] recently introduced an efficient label selection paradigm, which selects a sparse set of uncertain pixels for human labelling to train semantic segmentation models. Benenson and Ferrari [22] show that selecting sparse sets of to-be-labelled pixels reduces human labelling effort compared to dense pixel-wise human labels. Similarly, Xie et al. [19] propose a to-be-labelled pixel or region selection criterion for domain shift scenarios. In contrast to previous robotic planning methods [12–14], which rely on dense pixel-wise human labels, our work utilises a new sparse human label selection strategy inspired by Xie et al. [19] to drastically reduce human labelling effort.

**Semi-supervised semantic segmentation** methods build upon a low budget of human-labelled training samples methods and improve model performance further by generating pseudo labels from model predictions of unlabelled data [23, 24]. Our work leverages a low number of sparsely human-labelled samples and combines them with automatically generated pseudo labels. In contrast to image-based pseudo label methods [23, 24], our approach renders pseudo labels from a semantic map in an uncertainty-aware fashion.

**Informative path planning** aims to maximise the collected information in initially unknown environments subject to robot constraints, such as mission time [4, 25, 26]. Traditional *non-adaptive* approaches pre-compute static paths while *adaptive* methods actively re-plan paths online based on collected measurements [4, 27]. Our work focuses on *adaptive* methods for active learning in semantic segmentation since they account for varying semantics and changing model uncertainties after network re-training.

Sampling-based techniques, such as receding-horizon planning for information gathering [27] or variants of Monte-Carlo tree search [3, 28, 29], solve the informative path planning problem in a computationally efficient way. Similarly, optimisation-based strategies exploit algorithms such as the covariance matrix adaptation evolution strategy [4, 7] to directly maximise objective functions. Geometric methods select potentially informative candidate robot poses at the frontiers of explored and unknown space [25, 26]. The above-mentioned works address informative path planning for classical exploration or monitoring tasks. In contrast, we develop a geometric planning approach to improve robot vision using semi-supervised active learning.

Recent works in active learning for semantic segmentation using robotic platforms follow either the paradigm of self-supervision without human labels [16, 17] or full human supervision for selected informative images requiring dense pixel-wise labels [12–14]. Zurbrügg et al. [16] fuse 2D semantic predictions of a pre-trained network into a 3D map to automatically generate semantic labels for continual network re-training. Similarly, Chaplot et al. [17] train a viewpoint selection policy with reinforcement learning in simulation to target uncertain map parts. Despite not relying on human labels, these self-supervised methods require large human-labelled in-domain pre-training datasets in indoor scenes to produce high-quality pseudo labels. Further, systematic prediction errors prevent learning specific semantics [17]. In contrast, Blum et al. [12] propose a local planner to collect pixel-wise human-labelled data with high training data novelty [30]. Rückin et al. [14] propose a general map-based planner for fully supervised active learning to improve semantic segmentation. While these works do not require large pre-training datasets, they still depend on dense pixel-wise human labels for model training.

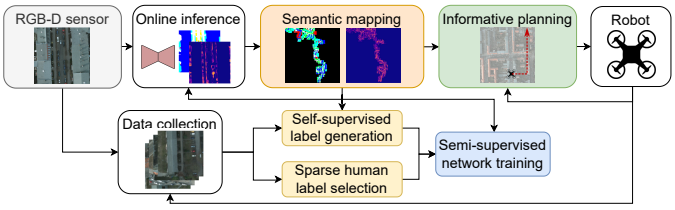


Fig 2: During a mission, a semantic segmentation network predicts pixel-wise semantics and model uncertainties from an RGB-D image. Both are fused into an uncertainty-aware semantic environment map (Sec. III-A). Our planner guides the collection of training data for network re-training based on the robot state and map belief (Sec. III-B). After a mission, the collected data is labelled using two sources of labels: (i) a human annotator labels a sparse set of informative pixels, and (ii) we automatically render pseudo labels from the semantic map in an uncertainty-aware fashion.

Our approach combines the advantages of self- and fully supervised approaches into a semi-supervised adaptive informative path planning framework. We maintain the general applicability of fully supervised approaches while reducing human labelling efforts in active learning for robotic vision.

## III. Our Approach

We present an adaptive informative path planning framework for semi-supervised active learning in semantic segmentation. Considering a robot equipped with an RGB-D sensor, our goal is to collect images in an initially unknown environment to improve semantic perception with minimal human labelling effort. Fig. 2 summarises our framework. We predict pixel-wise semantics and associated model uncertainties to update a probabilistic semantic environment map (Sec. III-A). Based on the robot pose, flight budget, and map belief, we plan paths to adaptively collect training data in environment areas of high model uncertainty (Sec. III-B). After a mission, we select a sparse set of informative to-be-human-labelled pixels in the collected images (Sec. III-C). We combine them with pseudo labels automatically rendered from the online-built semantic map in an uncertainty-aware fashion for network re-training (Sec. III-D).

### 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 a probabilistic multi-layered semantic environment mapping to fuse geometric and semantic information. The environment is discretised into three voxel maps $\mathcal{M}_G : V \to \{0,1\}^{W\times L\times H}$, $\mathcal{M}_S : V \to \{0,1\}^{K\times W\times L\times H}$, $\mathcal{M}_U : V \to [0,1]^{W\times L\times H}$ defined over $W \times L \times H$ spatially independent voxels $V$. $\mathcal{M}_G$ captures the geometric occupancy information, $\mathcal{M}_S$ stores the semantics, and $\mathcal{M}_U$ stores the associated model uncertainties.

The semantic map $\mathcal{M}_S$ consists of $K$ layers with one layer per class. At each time step, a new RGB-D image arrives. The probabilistic semantic predictions and model uncertainties are inferred using a semantic segmentation model and Monte-Carlo dropout [13]. We project the depth image, semantic predictions, and model uncertainties into the environment using the intrinsics and extrinsics of the RGB-D sensor. The geometric map $\mathcal{M}_G$ and semantic map $\mathcal{M}_S$ are recursively updated by probabilistic occupancy grid mapping [31]. The model uncertainty map $\mathcal{M}_U$ is updated by maximum likelihood estimation.

Additionally, we maintain a count map $\mathcal{M}_T : V \to \mathbb{N}^{W\times L\times H}$ to track the occurrences in the human-labelled training data utilised in our planning objective. As the semantic segmentation model is re-trained after each robot mission, the semantic predictions and model uncertainties change. Following Rückin et al. [14], 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 informative planning.

### B. Adaptive Informative Path Planning

Our planner is designed to collect new training data in initially unknown environments given mission budget constraints. 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 planning method searches for a path $\psi^* = (\mathbf{p}_1, \dots, \mathbf{p}_N) \in \Psi$ with a variable number $N \in \mathbb{N}$ of robot poses $\mathbf{p}_i \in \mathbb{R}^D$, $i \in \{1, \dots, N\}$, in the set of potential paths $\Psi$, that maximises an information criterion $I : \Psi \to \mathbb{R}_{\geq 0}$:

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

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

At each time step $t$, we adaptively re-plan the next-best robot pose $\mathbf{p}^*_{t+1}$ to collect informative training data. We utilise a geometric frontier-based planner [14, 32] guided by the information criterion $I$. The information criterion estimates the effect of a candidate training image recorded at a robot pose on a semantic segmentation model's performance. Based on the geometric map belief $\mathcal{M}^t_G$, we assign each voxel $v \in V$ to one of three disjoint sets of voxels $V_F \cup V_U \cup V_O = V$ containing the free, unknown, and surface voxels, respectively. To generate potentially informative robot pose candidates $\mathbf{p}^c_{t+1} \in \mathbb{R}^D$, we equidistantly sample poses $\mathbf{p}^c_{t+1}$ along the frontiers of free and unknown space reachable within the remaining mission budget. A frontier is a set of connected free voxels in $V_F$ with neighbouring unknown voxels in $V_U$ [32]. The information value $I(\mathbf{p}^c_{t+1})$ of a candidate pose $\mathbf{p}^c_{t+1}$ is defined as [16]:

$$I(\mathbf{p}^c_{t+1}) = \sum_{v\in \mathrm{Img}(\mathbf{p}^c_{t+1})} \begin{cases} 0 & , \text{if } v \in V_F \\ c_u & , \text{if } v \in V_U \\ \frac{\mathcal{M}^t_U(v)}{\mathcal{M}^t_T(v)} & , \text{if } v \in V_O\,, \end{cases} \tag{2}$$

where $c_u \in \mathbb{R} \geq 0$ is a uniform model uncertainty prior fostering exploration of unobserved environment areas, and $\mathrm{Img}(\mathbf{p}^c_{t+1})$ is a rendered 2D image of voxels visible from $\mathbf{p}^c_{t+1}$ with resolution $w' \times h'$. We obtain $\mathrm{Img}(\mathbf{p}^c_{t+1})$ by ray casting into the geometric map belief $\mathcal{M}^t_G$ from pose $\mathbf{p}^c_{t+1}$. While casting a ray from $\mathbf{p}^c_{t+1}$, only free voxels are treated

as traversable. Unknown or occupied voxels are assumed to be reflective. Pixels corresponding to rays that traverse only free voxels are assigned zero information value as we assume semantics are only assigned to surfaces. If a surface voxel reflects a ray, its effect on semantic segmentation performance after model re-training is estimated by its model uncertainty. To trade-off between model uncertainty and training data diversity, we normalise a voxel's information value by its number of occurrences in the training dataset.

### C. Semi-Supervised Training

The main contribution of our approach is a semi-supervised training strategy for improving the robot's semantic perception. We utilise a semantic segmentation network $\mathbf{f}_\theta(\mathbf{z}) = p(\mathbf{y} \,|\, \mathbf{z}) \in [K \times w \times h]$ parameterised by $\theta$ to predict the pixel-wise probabilities of $K$-class semantic labels $\mathbf{y} \in \{1, \dots, K\}^{w \times h}$ given an input RGB image $\mathbf{z}$ of resolution $w \times h$. We follow Rückin et al. [13] to estimate pixel-wise model uncertainties $\mathbf{u} \in [0, 1]^{w \times h}$ via Monte-Carlo dropout [18]. To maximise model performance, we combine human labels $\mathbf{Y}_l = \{\mathbf{y}_l^1, \dots, \mathbf{y}_l^{N_l}\}$ of images $\mathbf{Z}_l = \{\mathbf{z}_l^1, \dots, \mathbf{z}_l^{N_l}\}$ with pseudo labels $\mathbf{Y}_u = \{\mathbf{y}_u^1, \dots, \mathbf{y}_u^{N_u}\}$ of images $\mathbf{Z}_u = \{\mathbf{z}_u^1, \dots, \mathbf{z}_u^{N_u}\}$, where $N_l$ and $N_u$ are the numbers of human-labelled and pseudo-labelled images. To reduce human labelling effort, we consider a sparse set of human-labelled pixels $\mathbf{y}_l^i$ and pseudo-labelled pixels $\mathbf{y}_u^i$. Each non-labelled pixel in some label $\mathbf{y}_l^i$ or $\mathbf{y}_u^i$ is assigned a void class $N^v \in \{1, \dots, K\}$. During training, we mask the loss with $\mathbb{I}_{\mathbf{y} \neq N^v} \in \{0, 1\}^{w \times h}$, where $\mathbb{I}_{\mathbf{y} \neq N^v}$ is zero for each pixel with class $N^v$. The model $f_\theta$ is trained to minimise the following cross-entropy loss function:

$$\begin{aligned}\mathcal{L}(\theta) = &\frac{1}{N_l \alpha} \sum_{i=1}^{N_l} -\log\left(\mathbf{f}_\theta(\mathbf{z}_l^i)^{(\mathbf{y}_l^i, :, :)}\right) \mathbb{I}_{\mathbf{y}_l^i \neq N^v} + \\ &\frac{1}{N_u \alpha} \sum_{i=1}^{N_u} -\log\left(\mathbf{f}_\theta(\mathbf{z}_u^i)^{(\mathbf{y}_u^i, :, :)}\right) \mathbb{I}_{\mathbf{y}_u^i \neq N^v}\end{aligned} \tag{3}$$

where $\alpha \in \mathbb{N}$ is the number of labelled pixels per image and $\mathbf{f}_\theta(\mathbf{z})^{(\mathbf{y}, :, :)}$ are the probabilities of ground truth semantics $\mathbf{y}$.

Combining ideas from Shin et al. [20] and Xie et al. [19], 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, for all newly collected images $\mathbf{z}_l$ recorded at planned poses maximising Eq. (2), for each pixel $(m, n)$, we predict semantic probabilities $p(\mathbf{y} \,|\, \mathbf{z}_l)^{(:, m, n)}$ and extract the maximum likelihood label $\tilde{\mathbf{y}}_l^{(m,n)} = \operatorname{argmax}_{k \in [K]} p(\mathbf{y} \,|\, \mathbf{z}_l)^{(k, m, n)}$. We compute each pixel's region impurity score [19] as follows:

$$\begin{aligned}R_r(\mathbf{z}_l)^{(m,n)} &= -\sum_{k=1}^{K} \log\left(\frac{|N_r^k(m,n)|}{(2r+1)^2}\right) \frac{|N_r^k(m,n)|}{(2r+1)^2}, \\ N_r^k(m,n) &= \left\{(i,j) \in N_r(m,n) \,|\, \tilde{\mathbf{y}}_l^{(i,j)} = k\right\},\end{aligned} \tag{4}$$

where $N_r(m, n) = \{(i, j) \,|\, |i - m| \leq r, |j - n| \leq r\}$ is the set of $r$-step neighboring pixels of $(m, n)$. Intuitively, the region impurity for human labelling a pixel is high whenever the number of different classes predicted within the pixel's $r$-step neighbourhood is high, as a well-trained model should predict locally non-cluttered semantics. In contrast to Xie et al. [19], we do not greedily select the $\alpha$ pixels per image that maximise region impurity. Instead, per image, we sample $\alpha$ pixels uniformly at random from the $\beta\,\%$ pixels with the highest region impurity to foster human label diversity. While $\alpha$ sets the user-defined human labelling budget, $\beta$ implicitly provides a lower bound for a pixel's information value. Experimentally, we found that smaller values $\beta \leq 10\,\%$ ensure informative pixel selection, while $\beta \to 100\,\%$ lead to inefficient random pixel selection. Further, both region impurity and random sampling are crucial for maximising model performance.

### D. Self-Supervised Pseudo Label Generation

Similarly to self-supervised robotic active learning approaches [15, 16], after a mission is finished, we utilise our incrementally online-built uncertainty-aware semantic map (Sec. III-A) to generate pseudo labels $\mathbf{Y}_u$ in a self-supervised fashion. We record to-be-pseudo-labelled images $\mathbf{z}_u \in \mathbf{Z}_u$ equidistantly between two poses planned for collecting to-be-human-labelled images (Eq. (2)) to maximise training data diversity. Given a robot pose $\mathbf{p}_u \in \mathbb{R}^D$ at which $\mathbf{z}_u$ is recorded, we render a pixel-wise probabilistic pseudo label $p(\mathbf{y}_u \,|\, \mathbf{p}_u, \mathcal{M}_S) \in [0, 1]^{K \times w \times h}$ from the semantic map belief $\mathcal{M}_S$ at the image resolution $w \times h$. Then, for each pixel $(m, n)$, we extract the maximum likelihood pseudo label $\mathbf{y}_u^{(m,n)} = \operatorname{argmax}_{k \in [K]} p(\mathbf{y}_u \,|\, \mathbf{p}_u, \mathcal{M}_S)^{(k,m,n)}$. Similarly, we render the corresponding pixel-wise model uncertainty $\mathbf{u}_u \in [0, 1]^{w \times h}$ from the model uncertainty map $\mathcal{M}_U$. If a ray corresponding to pixel $(m, n)$ is not reflected by a surface voxel in $\mathcal{M}_G$, we assign the void class $\mathbf{y}_u^{(m,n)} = N^v$.

In contrast to previous works [15, 16], we only use a sparse set of $\alpha$ pseudo-labelled pixels per image $\mathbf{z}_u$ to train the network via Eq. (3) to balance the human and pseudo label supervision. We extend the approach of Shin et al. [20] to a new pixel selection procedure for sparse pseudo labels $\mathbf{y}_u$ that trades off between semantic map uncertainty and pseudo label diversity. After each mission, for all images $\mathbf{z}_u$ collected in any of the previous missions, we (re-)render pseudo labels $\mathbf{y}_u$ and model uncertainties $\mathbf{u}_u$ based on the most recent map beliefs $\mathcal{M}_S$ and $\mathcal{M}_U$. Similar to the human-labelled pixel selection (Sec. III-C), for each image, we sample $\alpha$ pixels $(m, n)$ at random from the $\beta\%$ pixels with the lowest map-based model uncertainty $\mathbf{u}_u^{(m,n)}$. Non-sampled pixels are assigned the void class. We found that providing an implicit upper bound $\beta$ for model uncertainty yields higher semantic segmentation performance than random sampling as it acts as a proxy to the pseudo label quality. Further, $\beta \leq 10\,\%$ usually ensures moderate model performance improvements.

## IV. Experimental Results

Our experiments are designed to assess the performance of our approach. They support the claims made in this paper. First, we show that our method for selecting human-labelled pixels outperforms state-of-the-art pixel selection methods in our robotic planning context (Sec. IV-B). Second, we validate that combining our uncertainty-aware pseudo labels with human labels improves semantic segmentation performance

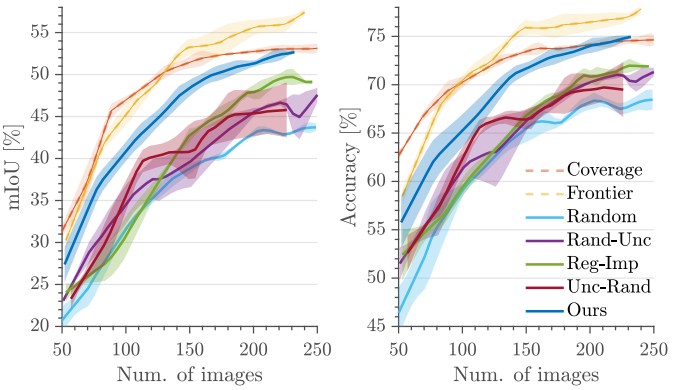


Fig 3: Comparison of label selection methods with $\alpha = 1000$ human-labelled pixels per image using our frontier planner on ISPRS Potsdam. Frontier (yellow) and coverage (orange) planners use densely labelled images indicating performance upper bounds. Results are averaged over three runs. Shaded regions indicate one standard deviation. Our proposed method (dark blue) outperforms state-of-the-art pixel selection methods.

and drastically reduces the number of human-labelled pixels compared to fully supervised approaches while maintaining similar performance (Sec. IV-C). Third, our semi-supervised active learning approach outperforms self-supervised active learning approaches (Sec. IV-D).

### A. Experimental Setup

**Baseline & Dataset.** We compare our frontier planner against a coverage-based strategy that pre-computes paths to maximise spatial coverage [14]. We evaluate our approach on the real-world 7-class ISPRS Potsdam orthomosaic dataset [33] and simulate 10 subsequent unmanned aerial vehicle (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 $400\,\text{px} \times 400\,\text{px}$ [14].

**Evaluation Metrics.** We evaluate semantic segmentation performance (dependent variable) over the number of human-labelled training images or pixels (independent variable). We use mean Intersection-over-Union (mIoU) [34] and pixel-wise accuracy [35] to quantify semantic segmentation performance. We run three trials per experiment and report the mean and standard deviation performance curves.

**Implementation Details.** We use Bayesian ERFNet [13] pre-trained on the Cityscapes dataset [34]. Re-training after each mission starts from this checkpoint. The model is trained until convergence on the validation set. We use a one-cycle learning rate scheduler, a batch size of 8, and weight decay $\lambda = (1-p)/2N$, where $p = 0.5$ is the dropout probability, and $N = N_l + N_u$ is the number of training images [9]. The human and pseudo label pixel selection lower bounds are $\beta = 5\%$, and the $r$-neighborhood of the human label selection criterion is set to $r = 1$. In practice, our approach allows for using any user-defined model.

### B. Targeted Human Label Selection

The first set of experiments shows that targeted human label selection improves semantic segmentation performance and reduces human labelling effort. We verify that our method (i) outperforms state-of-the-art pixel selection methods in the robotic planning context and (ii) improves semantic segmentation performance over non-targeted pixel selection with higher gains for lower human labelling budgets. The experiments are conducted using human labels only.

We compare our human-labelled pixel selection method (*Ours*, Sec. III-C) against four pixel selection methods for a low human labelling budget of $\alpha = 1000 \approx 0.6\%$ pixels for each collected image by our frontier planner (Sec. III-B). Namely: (i) sample $\alpha$ pixels from the $\beta\%$ most uncertain pixels [20] (*Unc-Rand*); (ii) sample $\beta\%$ pixels at random, then select the $\alpha$ most uncertain pixels [20] (*Rand-Unc*); (iii) select $\alpha$ pixels uniformly at random (*Random*); and (iv) select $\alpha$ pixels with the highest region impurity in an $r$-

TABLE I: Per-class IoU comparison of sparse label selection methods with $\alpha = 1000$ human-labelled pixels per image using our frontier planner on ISPRS Potsdam. *Dense* uses dense pixel-wise human-labelled images indicating the performance upper bound.

| Method | Mission | Surface | Building | Vegetation | Tree | Car | Clutter |
|---|---|---|---|---|---|---|---|
| Random | 3 | 53.98 | 51.00 | 40.58 | 20.87 | 28.54 | 7.58 |
| Unc-Rand | | 58.93 | 60.89 | 43.09 | 25.15 | 42.04 | 11.42 |
| Reg-Imp | | 51.17 | 48.84 | 39.96 | 15.29 | 0.00 | 8.26 |
| **Ours** | | **59.47** | **65.74** | **46.37** | **33.74** | **47.20** | **15.36** |
| Dense | | 63.93 | 70.39 | 49.46 | 35.82 | 60.95 | 10.40 |
| Random | 6 | 59.16 | 63.30 | 43.33 | 31.62 | 44.68 | 11.63 |
| Unc-Rand | | 61.87 | 68.99 | 42.50 | 29.80 | 52.57 | **16.60** |
| Reg-Imp | | 60.19 | 69.61 | 46.68 | 30.83 | 59.49 | 12.59 |
| **Ours** | | **65.99** | **72.83** | **51.56** | **41.16** | **61.07** | 15.53 |
| Dense | | 71.08 | 77.72 | 53.14 | 45.80 | 68.81 | 17.56 |
| Random | 9 | 59.38 | 64.71 | 43.80 | 33.21 | 50.54 | 11.33 |
| Unc-Rand | | 62.40 | 70.19 | 46.68 | 30.91 | 57.32 | 14.92 |
| Reg-Imp | | 62.31 | 71.78 | 46.87 | 36.92 | 64.57 | 12.33 |
| **Ours** | | **67.94** | **74.54** | **52.00** | **43.37** | **66.50** | **16.67** |
| Dense | | 71.23 | 78.60 | 52.79 | 48.52 | 71.57 | 20.11 |

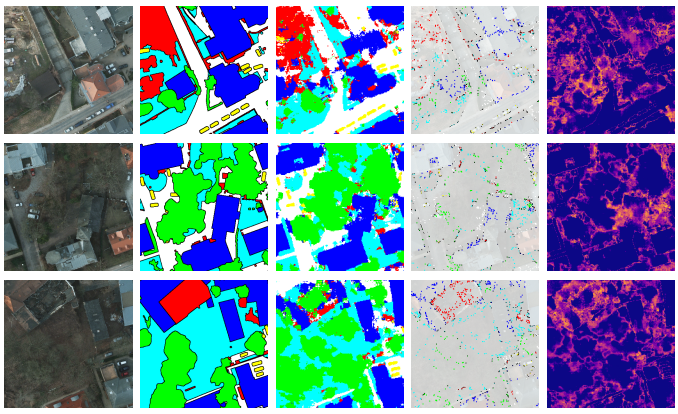


Fig 4: Qualitative results of our human label pixel selection method on ISPRS Potsdam. Columns from left to right: RGB input, ground truth, prediction, pixels selected for re-training, model uncertainty. Selected pixels are expanded to their one-pixel neighbourhood for visualisation. Our method selects pixels in areas of cluttered predictions, often corresponding to misclassified regions.

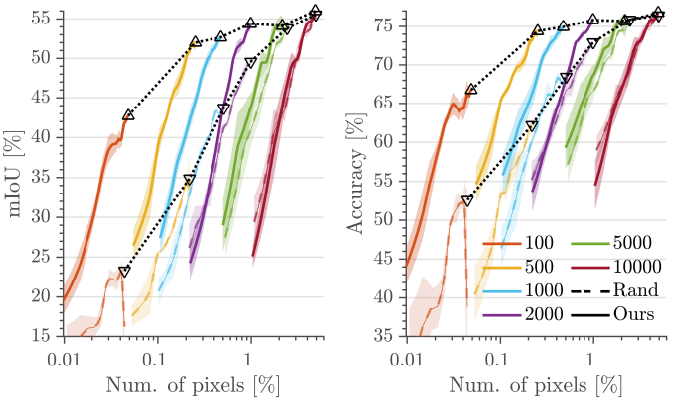


Fig 5: Comparison of our human label selection method (solid lines) to random label selection (dashed transparent lines) over varying labelling budgets $\alpha \in [100, 10000]$ px using our frontier planner on ISPRS Potsdam. Results are averaged over three runs. Shaded regions indicate one standard deviation. The performance gain of our method drastically increases for lower labelling budgets.

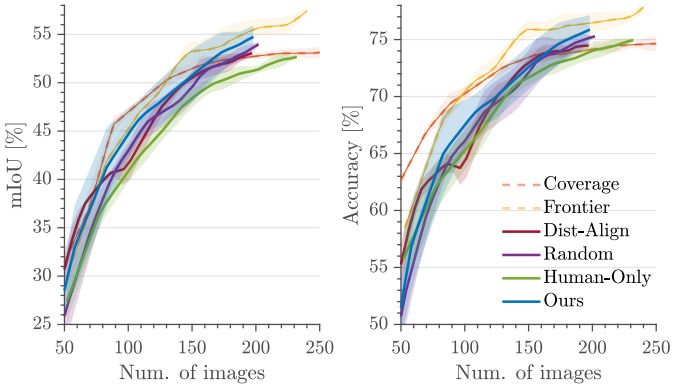


Fig 6: Comparison of pseudo label selection methods with $\alpha = 1000$ human- and pseudo-labelled pixels per image using our frontier planner on ISPRS Potsdam. Frontier (yellow) and coverage (orange) planners use densely labelled images indicating performance upper bounds. Results are averaged over three runs. Shaded regions indicate one standard deviation. Our method (dark blue) outperforms other pseudo label selection methods.

neighborhood [19] (*Reg-Imp*), where $r = 1$ yields the best results. Additionally, we show results for the *Frontier* and *Coverage* planner utilising pixel-wise densely human-labelled images [14] as an upper performance bound.

Fig. 3 summarises the semantic segmentation performance of the different pixel selection methods. In line with previous fully supervised approaches [12, 14], the *Frontier* planner (yellow) using densely labelled images achieves the highest performance outperforming the non-adaptive *Coverage* planner (orange). Notably, our method (dark blue) shows the fastest improvement and highest final mIoU of $\approx 52.5\%$ of all pixel selection methods, significantly outperforming the second-best *Reg-Imp* method (green) reaching $\approx 49\%$ final mIoU. Particularly, our human label selection matches the final performance of the *Coverage* planner using only $\approx 0.6\%$ of the labelled pixels. Table I shows the superior per-class performance of our human label selection method, verifying its ability to select sparse but informative human labels for different semantics. Fig. 4 displays images collected during a mission, onboard semantic predictions, and corresponding human-labelled pixels selected with our method.

Fig. 5 shows the semantic segmentation performance of our targeted pixel selection method (solid lines) compared to randomly selecting human-labelled pixels (dashed transparent lines) over varying human labelling budgets. Noticeably, for budgets $\alpha \leq 2000\,\text{px} \approx 1.3\,\%$, our pixel selection method clearly outperforms random pixel selection. Favourably, the performance gain of our pixel selection method over random pixel selection drastically increases with lower human labelling budgets. For an extremely low budget of $\alpha = 100\,\text{px} \approx 0.06\,\%$, our targeted pixel selection method leads to a high final performance gain of $\approx 20\,\%$ mIoU.

### C. Uncertainty-Aware Pseudo Label Generation

The second set of experiments shows that uncertainty-aware generation of pseudo labels improves semantic segmentation performance. We validate that (i) our pseudo label selection method outperforms other selection strategies, (ii) combining

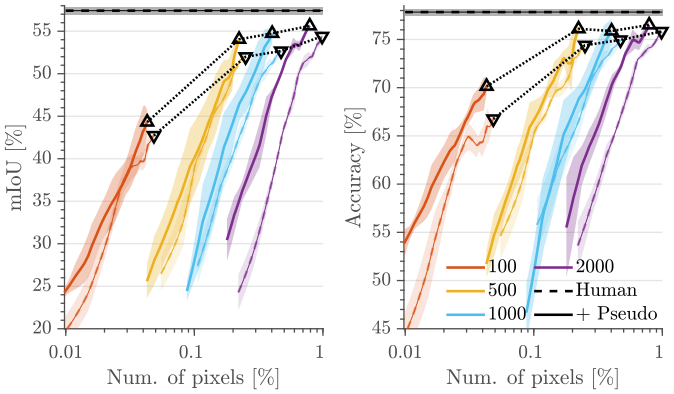


Fig 7: Comparison of our human label selection only (dashed transparent lines), and combined with our pseudo label selection (solid lines) over varying labelling budgets $\alpha \in [100, 2000]$ px per image using our frontier planner on ISPRS Potsdam. Results are averaged over three runs. Shaded regions indicate one standard deviation. Using our pseudo labels consistently improves performance.

our human label selection with our pseudo label selection consistently improves semantic segmentation performance across varying labelling budgets, and (iii) our semi-supervised approach drastically reduces the number of human-labelled pixels compared to fully supervised approaches while maintaining similar performance. The experiments are conducted using our human label selection method.

We compare our uncertainty-aware pseudo label selection (*Ours*, Sec. III-D) against two other pseudo label selection methods for a low human labelling budget of $\alpha = 1000\,\text{px} \approx 0.6\,\%$ per image. Namely: (i) we re-distribute the pseudo labels' class distribution to the true class distribution estimated by the human labels using per-class model uncertainty thresholds to select on average $\alpha$ pixels per image [23] (*Dist-Align*), and (ii) we randomly select $\alpha$ pixels per image (*Random*). We compare against using $\alpha$ human-labelled pixels per image only (*Human-Only*) and using the *Frontier* or *Coverage* planners leveraging dense human labels.

Fig. 6 summarises the performance of the different methods.

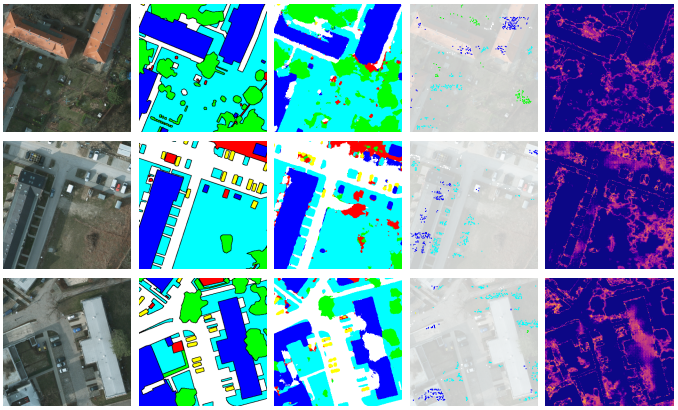


Fig 8: Qualitative results of our pseudo label generation on ISPRS Potsdam. Left to right: RGB input, ground truth, pseudo label, pixels selected for re-training, model uncertainty. Selected pixels are expanded to their one-pixel neighbourhood for visualisation. Our method selects low-uncertainty pixels to minimise label errors.

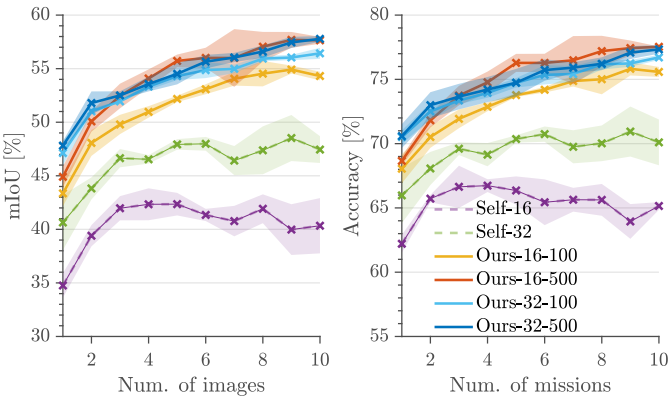


Fig 9: Comparison of our semi-supervised (solid lines) and a self-supervised approach (dashed transparent lines) with varying numbers of human-labelled pixels during deployment and densely human-labelled images for pre-training. Results are averaged over three runs on ISPRS Potsdam. Shaded regions indicate one standard deviation. Our semi-supervised approach outperforms the self-supervised approach for all labelling budget configurations.

Combining human labels with pseudo labels improves performance over sparse human labels only (green). This verifies our concept of leveraging both sparse human supervision and self-supervised pseudo labels to maximise performance. Our uncertainty-aware pseudo label selection method (dark blue) achieves $\approx 1-2\%$ mIoU higher than other methods (purple, red). Particularly, our semi-supervised approach outperforms the *Coverage* planner using only $\approx 0.6\,\%$ of the human-labelled pixels. Fig. 8 shows qualitative results for our method after mission completion.

Fig. 7 shows the performance using our human-labelled pixel selection method only (dashed transparent lines) and combining it with our uncertainty-aware pseudo labels (solid lines) over varying human labelling budgets $\alpha \in [0.06, 1.25]\,\%$ per image using our frontier planner. Combining sparse human and pseudo labels consistently improves performance by $\approx 2-3\,\%$ mIoU across varying budgets. This validates the superior performance of our semi-supervised approach over using sparse human labels only. Further, our semi-supervised approach rapidly closes the final performance gap to the fully supervised frontier planner proposed by Rückin et al. [14] (dashed black line). The fully supervised approach reaches a maximum performance of $\approx 57.5\%$ mIoU while our semi-supervised approach reaches $\approx 56\%$ mIoU with only $\approx 0.6\%$ of the human-labelled pixels. This shows that our semi-supervised approach requires two magnitudes fewer human-labelled pixels while reaching performance similar to previous fully supervised approaches.

### *D. Semi- vs. Self-Supervised Robotic Active Learning*

The third set of experiments shows that our semi-supervised active learning framework outperforms self-supervised approaches by a large margin under varying human labelling budgets for model pre-training and model re-training in the unknown environment.

Similar to self-supervised approaches for robotic continual learning and domain adaptation [15, 16], we utilise our frontier planner to guide uncertainty-driven training data collection and exploit the online-built map to generate dense pseudo labels. Current self-supervised approaches only work with pre-trained semantic segmentation models deployed in similar environments [15–17]. Although our semi-supervised method works in a completely unknown environment (Sec. IV-C), for comparing to self-supervised methods, we relax these assumptions and consider small amounts of densely human-labelled pre-training data randomly sampled from the deployment environment. Each approach starts with the same model checkpoint trained on the sampled pre-training data. Similar to the experience replay method of self-supervised approaches [15, 16], to achieve performance improvements in the self-supervised approach, the human-labelled pre-training data is additionally used for model re-training after a mission is completed.

Fig. 9 shows the semantic segmentation performance of our semi-supervised approach (solid lines) compared to the self-supervised approach (dashed lines) on ISPRS Potsdam with varying numbers of human-labelled pre-training data $\{16, 32\}$. For all human labelling budgets $\alpha \in \{100, 500\} \approx \{0.06, 0.3\}\%$ and all pre-training data budgets, our semi-supervised active learning approach outperforms self-supervision by a large margin. With a small number of 16 pre-training images and little human supervision of $\alpha = 100$ during the missions, our semi-supervised approach achieves higher final performance than the self-supervised approach with 32 pre-training images. Further, irrespective of the number of pre-training images, self-supervision fails to improve its performance after five missions. This suggests that semi-supervised active learning is necessary for maximally improving semantic segmentation in unknown or partially known environments. Although self-supervision benefits from minimal labelling requirements during deployment, it is inherently limited by its lack of knowledge and systematic prediction errors in unknown environments [17].

## V. Conclusions and Future Work

We proposed a novel adaptive informative path planning approach for semi-supervised active learning in robotic semantic perception with minimal human labelling effort. Our main contribution is a method for selecting sparse sets of informative pixels for human labelling and combining them with automatically generated pseudo labels rendered from an online-built uncertainty-aware semantic map. Our experimental results show that our sparse human-labelled pixel selection method outperforms state-of-the-art pixel selection methods. Combining human labels with pseudo labels further improves performance. Our semi-supervised approach drastically reduces human labelling effort compared to fully supervised methods while preserving similar performance and outperforming purely self-supervised approaches. Despite those encouraging results, future work could develop new methods to generate human labelling queries, e.g. by utilising foundation models [36], and automatically extract pseudo labels.

## 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, S. Sanner *et al.*, "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, pp. 1183–1192.

[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] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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)*, pp. 13 086–13 098, 2021.

[18] 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.

[19] 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.

[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] 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, pp. 9368–9377.

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

[23] R. He, J. Yang, and X. Qi, "Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2021, pp. 6930–6940.

[24] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[25] M. Ghaffari Jadidi, J. Valls Miro, and G. Dissanayake, "Gaussian Processes Autonomous Mapping and Exploration for Range-Sensing Mobile Robots," *Autonomous Robots*, vol. 42, no. 2, pp. 273–290, 2018.

[26] F. Chen, J. D. Martin, Y. Huang, J. Wang, and B. Englot, "Autonomous Exploration Under Uncertainty via Deep Reinforcement Learning on Graphs," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2020.

[27] G. A. Hollinger and G. S. Sukhatme, "Sampling-based Robotic Information Gathering Algorithms," *Int. Journal of Robotics Research (IJRR)*, vol. 33, no. 9, pp. 1271–1287, 2014.

[28] J. Ott, E. Balaban, and M. J. Kochenderfer, "Sequential Bayesian Optimization for Adaptive Informative Path Planning with Multimodal Sensing," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2023.

[29] J. Rückin, L. Jin, and M. Popović, "Adaptive informative path planning using deep reinforcement learning for UAV-based active sensing," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2022.

[30] N. Papernot and P. McDaniel, "Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning," *arXiv preprint arXiv:1803.04765*, 2018.

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

[32] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proc. of Int. Symp. on Computational Intelligence in Robotics and Automation*, 1997.

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

[34] 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.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[36] 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, pp. 4015–4026.