

Embodied Visual Active Learning for Semantic Segmentation

David Nilsson^{1,2*}, Aleksis Pirinen¹, Erik Gärtner^{1,2*}, Cristian Sminchisescu^{1,2}

¹Department of Mathematics, Faculty of Engineering, Lund University

²Google Research

{david.nilsson, aleksis.pirinen, erik.gartner, cristian.sminchisescu}@math.lth.se

Abstract

We study the task of *embodied visual active learning*, where an agent is set to explore a 3d environment with the goal to acquire visual scene understanding by actively selecting views for which to request annotation. While accurate on some benchmarks, today’s deep visual recognition pipelines tend to not generalize well in certain real-world scenarios, or for unusual viewpoints. Robotic perception, in turn, requires the capability to refine the recognition capabilities for the conditions where the mobile system operates, including cluttered indoor environments or poor illumination. This motivates the proposed task, where an agent is placed in a novel environment with the objective of improving its visual recognition capability. To study embodied visual active learning, we develop a battery of agents – both learnt and pre-specified – and with different levels of knowledge of the environment. The agents are equipped with a semantic segmentation network and seek to acquire informative views, move and explore in order to propagate annotations in the neighbourhood of those views, then refine the underlying segmentation network by online retraining. The trainable method uses deep reinforcement learning with a reward function that balances two competing objectives: *task performance*, represented as *visual recognition accuracy*, which requires exploring the environment, and the necessary *amount of annotated data* requested during active exploration. We extensively evaluate the proposed models using the photorealistic Matterport3D simulator and show that a fully learnt method outperforms comparable pre-specified counterparts, even when requesting fewer annotations.

1 Introduction

Imagine a household robot in a home it has never been before and equipped with a visual sensing module to perceive its environment and localize objects. If the robot fails to recognize some objects, or to adapt to changes in the environment, over time, it may not be able to properly perform its tasks. Much of the recent success of visual perception has been achieved by deep CNNs, e.g. in image classification (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014; He et al. 2016), semantic segmentation (Long, Shelhamer, and Darrell 2015; Chen et al. 2017)

and object detection (Ren et al. 2015; Redmon et al. 2016). Such systems may however be challenged by unusual viewpoints or domains, as noted e.g. by Ammirato et al. (2017) and Yang et al. (2019b). Moreover, a mobile household robot should ideally operate with lightweight, re-trainable and task-specific perception models, rather than large and comprehensive ones, which could be demanding computationally and not tailored to the needs of a specific house.

In practice, even in closed but large environments, developing robust scene understanding by exhaustive approaches may be difficult, as looking everywhere requires an excessive amount of annotation labor. All views are however not equally informative, as a view containing many diverse objects is likely more useful than one covering a single semantic class, e.g. a wall. This suggests that in learning visual perception one does not have to label exhaustively. As new, potentially difficult arrangements appear in an evolving environment, it would be useful to identify those automatically, based on the task and demand, rather than programmatically, by periodically re-training a complete model. Moreover, the agent could make the most out of its embodiment by propagating a given ground truth annotation using motion – as measured by the perceived optical flow – in that neighborhood. The agent can then self-train, online, for increased performance. The key questions are how should one explore the environment, how to select the most informative views to annotate, and how to make the most out of them. We analyze these questions in an *embodied visual active learning* framework, illustrated in fig. 1.

To ground the embodied visual active learning task, in this work we measure visual perception ability as semantic segmentation accuracy. The agent is equipped with a semantic segmentation system and must move around and request annotations in order to refine it. After exploring the scene the agent should be able to accurately segment all views in the explored area. This requires an exploration policy covering different objects from diverse viewpoints and selecting sufficiently many annotations to train the perception model. The agent can also propagate annotations to different nearby viewpoints using optical flow and then self-train. We develop a battery of methods, ranging from pre-specified ones to a fully trainable deep reinforcement learning-based agent, which we evaluate extensively in the photorealistic Matterport3D environment (Chang et al. 2017).

*Work was partially performed during a Google internship.
Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

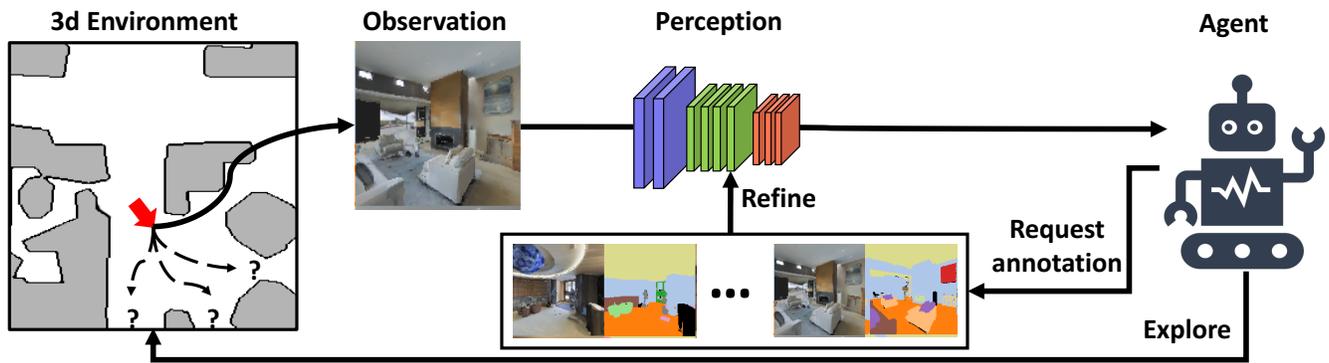


Figure 1: Embodied visual active learning. An agent in a 3d environment must explore and occasionally request annotation in order to efficiently refine its visual perception. The navigation component makes this task significantly more complex than traditional active learning, where the data pool over which the agent queries annotations, either in the form of image collections or pre-recorded video streams, is static and given.

In summary, our main contributions are:

- We study the task of *embodied visual active learning*, where an agent should explore a 3d environment to acquire visual scene understanding by actively selecting views for which to request annotation. The agent then propagates information by moving in the neighborhood of those views and self-trains;
- In our setup, visual learning and exploration can inform and guide one another since the recognition system is selectively and gradually refined during exploration, instead of being trained at the end of a trajectory on a full set of densely annotated views;
- We develop a variety of methods, both learnt and pre-specified, to tackle our task in the context of semantic segmentation;
- We perform extensive evaluation in a photorealistic 3d environment and show that a fully learnt method outperforms comparable pre-specified ones.

2 Related Work

The embodied visual active learning setup leverages several computer vision and machine learning concepts, such as embodied navigation, active learning and active vision. There is substantial recent literature on embodied agents navigating in real or simulated 3d environments, especially given the recent emergence of large-scale simulators (Savva et al. 2019; Kolve et al. 2017; Xia et al. 2018; Dosovitskiy et al. 2017; Savva et al. 2017).

We here briefly review variants of embodied learning. In Embodied Question Answering (Das et al. 2018; Wijmans et al. 2019; Yu et al. 2019), an agent is given a question, e.g. "What color is the car?". The agent must typically explore the environment quite extensively in order to be able to answer. Zhu et al. (2017); Mousavian et al. (2019) task the agent with reaching a target view using as few steps as possible. The agent receives the current view and the target as inputs in each step. In point-goal navigation (Mishkin, Dosovitskiy, and Koltun 2019; Sax et al. 2019; Savva et al.

2019; Gupta et al. 2017) the agent is given coordinates of a target to reach using visual information and ego-motion. In visual exploration (Ramakrishnan, Jayaraman, and Grauman 2020; Fang et al. 2019; Chen, Gupta, and Gupta 2019; Qi et al. 2020; Zheng et al. 2019; Chaplot et al. 2020) the task is to explore an unknown environment as quickly as possible, by covering the whole scene area. In Ammirato et al. (2017); Yang et al. (2019b) an agent is tasked to navigate an environment to increase the accuracy of a pre-trained recognition model, e.g. by moving around occluded objects. This is in contrast to our work where the goal is to collect views for *training* a perception model. Whereas in Ammirato et al. (2017); Yang et al. (2019b) the agent is spawned close to the target object, we cannot make such assumptions, as our task is not only to accurately recognize a single object or view, but to do so for *all* views in the potentially large area explored by the agent.

There are relations to curiosity-driven learning (Pathak et al. 2017; Yang et al. 2019a), in that we also seek an agent which visits novel views (states). In Pathak et al. (2017), exploration is aided by giving rewards based on the prediction error of a self-supervised inverse-dynamics model. This is a task-independent exploration strategy useful to search 2d or 3d environments during training. In our setup, exploration is task-specific in that it is aimed specifically at refining a visual recognition system in a novel environment. Moreover, we use semi-dense rewards for both visual learning and for exploration. Hence we are not operating using sparse rewards where curiosity approaches often outperform other methods.

Our work is also related to Song, Zhang, and Xiao (2015); Pot, Toshev, and Kosecka (2018); Zhong et al. (2018); Wang et al. (2019). Differently from us, Song, Zhang, and Xiao (2015) uses hand-crafted annotation and exploration strategies, aiming to label all voxels in a 3d reconstruction by selecting a subset of frames covering all voxels. This is a form of exhaustive annotation and a visual perception system is not trained. Hence the system can only analyze objects in annotated voxels. In our setup the agent is instead tasked with both exploration and the selection of views to

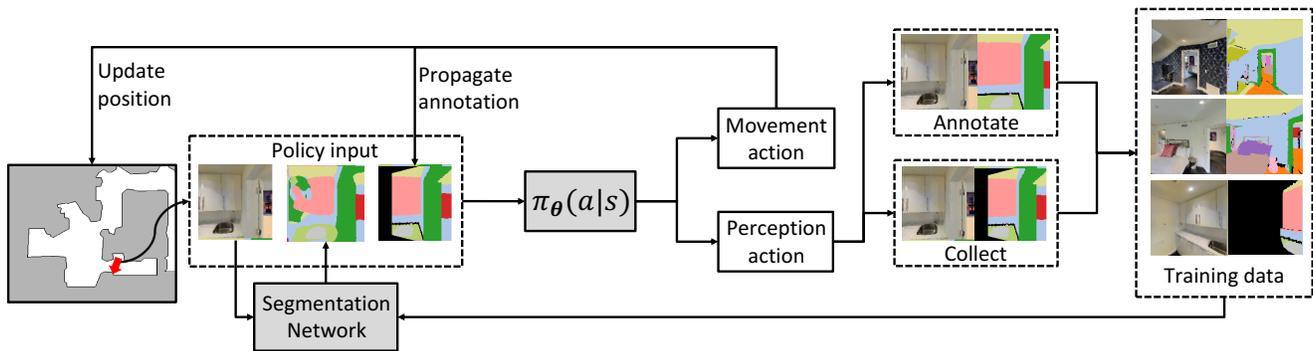


Figure 2: Embodied visual active learning for semantic segmentation. A first-person agent is placed in a room and a deep network predicts the semantic segmentation of the agent’s view. Based on the view and its segmentation, the agent can either select a *movement action* to change position and viewpoint, or select a *perception action* (*Annotate* or *Collect*). *Annotate* adds the current view and its ground truth segmentation to the pool of training data for the segmentation network, while *Collect* is a cheaper version (no additional supervision required) where the current view and the last annotated view – propagated to the agent’s current position using optical flow – is added to the training set. The propagated annotation is also a policy input for the learnt agent in §3.3. After a perception action, the segmentation network is refined on the current training set. The embodied visual active learning process is considered successful if, after selecting a limited number of *Annotate* actions or an exploration budget is exhausted, the segmentation network can accurately segment any other view in the environment where the agent operates. Note that the map (left) is *not* provided as input to the learnt agent in §3.3.

annotate, and we learn a perception module aiming to generalize to unseen views. In contrast to us, Pot, Toshev, and Kosecka (2018); Zhong et al. (2018); Wang et al. (2019) do not consider an agent choosing where to move in the environment, nor which parts to label. Instead, they use all views seen when following a pre-specified path for training a visual recognition system. Pot, Toshev, and Kosecka (2018) use an object detector obtained by self-supervised learning and clustering. Zhong et al. (2018); Wang et al. (2019) use constraints from SLAM to improve a given segmentation model. This approach could in principle complement our label propagation, and is orthogonal to our main proposals.

Next-best-view (NBV) prediction (Jayaraman and Grauman 2018; Xiong and Grauman 2018; Johns, Leutenegger, and Davison 2016; Jayaraman and Grauman 2016; Song et al. 2018; Gärtner, Pirinen, and Sminchisescu 2020) is superficially similar to our task. In Jayaraman and Grauman (2018) an agent is trained to reveal parts of a panorama and a model is built to complete all views of the panorama. Our setup allows free movement in an environment, hence it features a navigation component which makes our task more comprehensive. While NBV typically integrates information from all predicted views, our task requires the adaptive selection of only a subset of the views encountered during the agent’s navigation trajectory.

Active learning (Settles 2009; Fang, Li, and Cohn 2017; Lughofer 2012; Woodward and Finn 2017; Pardo et al. 2019; Feng et al. 2019) can be seen as the static version of our setup, as it considers approaches for learning what parts of a larger *pre-existing* and *static* training set should be fed into the training procedure, and in what order. We instead consider the active learning problem in an embodied setup, where an agent can move and actively select views for which to request annotation. Embodiment makes it possible to use

motion to propagate annotations, hence effectively generate new ones at no additional annotation cost. In essence, our work lays groundwork towards marrying the active vision and the active learning paradigms.

3 Embodied Visual Active Learning

Embodied visual active learning is an interplay between a first-person agent, a 3d environment and a trainable perception module. See fig. 1 for a high-level abstraction and fig. 2 for details of the particular task considered in this paper. The perception module processes images (views) observed by the agent in the environment. The agent can request annotations for views in order to refine the perception module. It should ideally request very few annotations as these are costly. The agent can also generate more annotations for free by neighborhood exploration using label propagation, such that when trained on that data the perception module becomes increasingly more accurate in the explored environment. To assess how successful an agent is on the task, we test how accurate the perception module is on multiple random viewpoints selected uniformly in the area explored by the agent.

Task overview. The agent begins each episode randomly positioned and rotated in a 3d environment, with a randomly initialized semantic segmentation network. The ground truth segmentation mask for the first view is given for the initial training of the segmentation network. The agent can choose *movement actions* (*MoveForward*, *MoveLeft*, *MoveRight*, *RotateLeft*, *RotateRight* with 25 cm movements and 15 degree rotations), or *perception actions* (*Annotate*, *Collect*). If the agent moves or rotates, the ground truth mask is propagated using optical flow. At any time, the agent may choose to insert the propagated annota-

tion into its training set with the `Collect` action, or to ask for a new ground truth mask with the `Annotate` action. After an `Annotate` action the propagated annotation mask is re-initialized to the ground truth annotation. After each perception action, the segmentation network S is refined on the training set, which is expanded with the new data point.

The agent’s performance is evaluated at the end of the episode. The goal is to maximize the mIoU and mean accuracy of the segmentation network on the views in the area explored by the agent. Specifically, a set of *reference views* are randomly sampled within a disc of radius r centered at the starting location, and the segmentation network is evaluated on these. Hence to perform well the agent is required to explore its surroundings, and it should refine its perception module in regions of high uncertainty.

3.1 Methods for the Proposed Task

We develop several methods to evaluate and study the embodied visual active learning task. All methods except the RL-agent issue the `Collect` action when 30% of the propagated labels are unknown and `Annotate` when 85% are unknown. The intuition is that the pre-specified methods should request annotation when most pixels are unlabeled. The specific hyperparameters of all models were set based on a validation set.

Random. Uniformly selects random movement actions. This baseline is thus a lower bound in terms of embodied exploration for this task.

Rotate. Continually rotates left. This method is useful in comparing with trainable agents that move and explore, i.e. to monitor what improvements can be expected from embodiment.

Bounce. Explores by walking straight forward until it hits a wall, then samples a new random direction and moves forward until it collides with a new wall, and so on. This agent quickly explores the environment.

Frontier exploration. This method builds a map, online, by using using depth and motion from the simulator (Yamauchi 1997). All pixels with depth within a 4m threshold are back-projected in 3d and then classified as either obstacles or navigable, based on height relative to the ground plane. This agent is confined to move within the reference view radius r , which is a choice to its advantage¹ as annotated views will more likely be similar to reference views that reside within that same radius.

Space filler. Follows a shortest space filling curve within the reference view radius r , and as r increases the entire environment is explored. This baseline makes strong and somewhat less general (or depending on the application, altogether unrealistic) assumptions in order to create a path: knowing the floor plan in advance, as well as which

¹This ensures it is evaluated under ideal conditions in contrast to the RL-agent in §3.3.

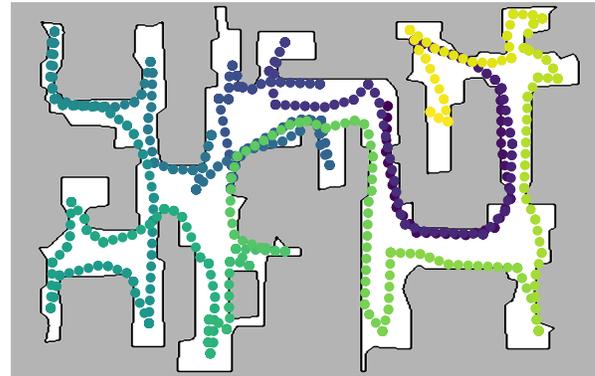


Figure 3: An example of a space filling curve in a Matterport3D floor plan. Methods based on the space filler assume complete spatial knowledge of the environment.

locations are reachable from the start. It also only moves within the reference view radius, and knows the shortest geodesic paths to take on the curve. Hence, this method can be considered an upper bound for other methods. The space filling curve is computed by placing a grid of nodes onto the floor plan (1m resolution, using a sampling and reachability heuristic), and then finding the shortest path around it with an approximate traveling salesman solver. Fig. 3 shows a space filling curve in a Matterport3D floor plan.

RL-agent. This fully trainable method we develop jointly learns exploration and perception actions in a reinforcement learning framework. See the full description in §3.3.

3.2 Semantic Segmentation Network

Each method uses the same FCN-inspired deep network (Long, Shelhamer, and Darrell 2015) for semantic segmentation. The network consists of 3 blocks of convolutional layers, each containing 3 convolutional layers with kernels of size 3×3 . The first convolutional layer in each block uses a stride of 2, which halves the resolution. For each block the number of channels doubles, using 64, 128 and 256 channels respectively. Multiple predictions are made using the final convolutional layers of each block. The multi-scale predictions are resized to the original image resolution using bilinear interpolation and are finally summed up, resulting in the final segmentation estimate. Note that we have deliberately chosen to make the network small so that it can be efficiently refined on new data.

At the beginning of each episode, the parameters are initialized randomly, and we train the network on the very first view, for which we always supply the ground truth segmentation. Each time `Annotate` or `Collect` is selected, we refine the network. Mini-batches of size 8, which always include the latest added labeled image, are used in training. We use random cropping and scaling for data augmentation. The network is refined either until it has trained for 1,000 iterations or until the accuracy of a mini-batch exceeds 95%. We use a standard cross-entropy loss averaged over all pixels. The segmentation network is trained using stochastic gradi-

ent descent with learning rate 0.01, weight decay 10^{-5} and momentum 0.9. To propagate semantic labels, we compute optical flow between consecutive viewpoints using PWC-Net (Sun et al. 2018). The optical flow is computed bidirectionally and only pixels where the difference between the forward and backward displacements is less than 2 pixels are propagated (Sundaram, Brox, and Keutzer 2010). We found that labels were reliably tracked over several frames when using 2 pixels as a threshold.

3.3 Reinforcement Learning Agent

To present the reinforcement-learning agent for our task, we begin with an explanation of the state-action representation and policy network, followed by the reward structure and finally policy training.

Actions, states and policy. The agent is represented as a deep stochastic policy $\pi_\theta(a_t|s_t)$ that samples an action a_t in state s_t at time t . The actions are MoveForward, MoveLeft, MoveRight, RotateLeft, RotateRight, Annotate and Collect. The full state is $s_t = \{\mathbf{I}_t, \mathbf{S}_t, \mathbf{P}_t, \mathbf{F}_t\}$ where $\mathbf{I}_t \in \mathbb{R}^{127 \times 127 \times 3}$ is the image, $\mathbf{S}_t = \mathcal{S}_t(\mathbf{I}_t) \in \mathbb{R}^{127 \times 127 \times 3}$ is the semantic segmentation mask predicted by the deep network \mathcal{S}_t (this network is refined over an episode; t indexes the network parameters at time t), $\mathbf{P}_t \in \mathbb{R}^{127 \times 127 \times 3}$ is the propagated annotation, and $\mathbf{F}_t \in \mathbb{R}^{7 \times 7 \times 2048}$ is a deep representation of \mathbf{I}_t (a ResNet-50 backbone feature map).

The policy consists of a *base processor*, a *recurrent module* and a *policy head*. The base processor consists of two learnable components: ϕ_{img} and ϕ_{res} . The 4-layer convolutional network ϕ_{img} takes as input the depth-wise concatenated triplet $\{\mathbf{I}_t, \mathbf{S}_t, \mathbf{P}_t\}$, producing $\phi_{img}(\mathbf{I}_t, \mathbf{S}_t, \mathbf{P}_t) \in \mathbb{R}^{512}$. Similarly, the 2-layer convolutional network ϕ_{res} yields an embedding $\phi_{res}(\mathbf{F}_t) \in \mathbb{R}^{512}$ of the ResNet features \mathbf{F}_t . An LSTM (Hochreiter and Schmidhuber 1997) with 256 cells constitutes the recurrent module, which takes as input $\phi_{img}(\mathbf{I}_t, \mathbf{S}_t, \mathbf{P}_t)$ and $\phi_{res}(\mathbf{F}_t)$. The input has length 1024. The hidden LSTM state is fed to the policy head, consisting of a fully-connected layer followed by a 7-way softmax which produces action probabilities.

Rewards. In training, the main reward is related to the mIoU improvement of the final segmentation network \mathcal{S}_T over the initial \mathcal{S}_0 on a reference set \mathcal{R} . The set \mathcal{R} is constructed at the beginning of each episode by randomly selecting views within a geodesic distance r from the agent’s starting location, and contains views with corresponding ground truth semantic segmentation masks. At the end of an episode of length T , the underlying perception module is evaluated on \mathcal{R} . Specifically, after an episode (with T steps), the agent receives as final reward:

$$R_T = \text{mIoU}(\mathcal{S}_T, \mathcal{R}) - \text{mIoU}(\mathcal{S}_0, \mathcal{R}) \quad (1)$$

To obtain a denser signal, tightly coupled with the final objective, we also give a reward proportional to the improvement of \mathcal{S} on the reference set \mathcal{R} after each Annotate

(*ann*) and Collect (*col*) action:

$$R_t^{ann} = \text{mIoU}(\mathcal{S}_t, \mathcal{R}) - \text{mIoU}(\mathcal{S}_{t-1}, \mathcal{R}) - \epsilon^{ann} \quad (2)$$

$$R_t^{col} = \text{mIoU}(\mathcal{S}_t, \mathcal{R}) - \text{mIoU}(\mathcal{S}_{t-1}, \mathcal{R}) \quad (3)$$

To ensure the agent does not request costly annotations too frequently, each Annotate action is penalized with a negative reward $-\epsilon^{ann}$ (we set $\epsilon^{ann} = 0.01$), as seen in (2). Such a penalty is *not* given for the free Collect action. Moreover, the dataset we use has 40 different semantic classes, but some are very rare and apply only to small objects, and some might not even be present in certain houses. We address this imbalance by computing the mIoU using only the 10 largest classes, ranked by the number of pixels in the set of reference views for the current episode.

While the rewards (1) - (3) should implicitly encourage the agent to explore the environment in order to request annotations for distinct, informative views, we empirically found useful to include an additional explicit exploration reward. Denote by $\{\mathbf{x}_i\}_{i=1}^{t-1} = \{(x_i, y_i)\}_{i=1}^{t-1}$ the positions the agent has visited up to time $t - 1$ in its current episode, and let $\mathbf{x}_t = (x_t, y_t)$ denote its current position. We define the exploration (*exp*) reward based on a kernel density estimate of the agent’s visited locations:

$$R_t^{exp} = a - bp_t(\mathbf{x}_t) := a - \frac{b}{t-1} \sum_{i=1}^{t-1} k(\mathbf{x}, \mathbf{x}_i) \quad (4)$$

where a and b are hyperparameters (both set to 0.003). Here $p_t(\mathbf{x}_t)$ is a Gaussian kernel estimate of the density with bandwidth 0.3m. It is large for previously visited positions and small for unvisited positions, thereby encouraging the agent’s expansion towards new places in the environment. The exploration reward is only given for movement actions. Note that the pose \mathbf{x}_i is only used to compute the reward R_t^{exp} and is not available to the policy via the state space.

Policy training. The policy network is trained using PPO (Schulman et al. 2017) based on the RLlib reinforcement learning package (Liang et al. 2018), as well as OpenAI Gym (Brockman et al. 2016). For optimization we use Adam (Kingma and Ba 2014) with batch size 512, learning rate 10^{-4} and discount rate 0.99. During training, each episode consists of 256 actions. The agent is trained for 4k episodes, which totals 1024k steps.

Our system is implemented in TensorFlow (Abadi et al. 2016), and it takes about 3 days to train an agent using 4 Nvidia Titan X GPUs. An episode of length 256 took on average about 3 minutes using a single GPU, and during training we used 4 workers with one GPU each, collecting rollouts independently. The runtime per episode varies depending on how frequently the agent decides to annotate, as training the segmentation network is the bottleneck and accounts for approximately 90% of the run-time. We used optical flow from the simulator to speed up policy training. For evaluation, the RL-agent and all other methods use PWC-Net to compute optical flow. The ResNet-50 feature extractor is pre-trained on ImageNet (Jia Deng et al. 2009) with weights frozen during policy training.

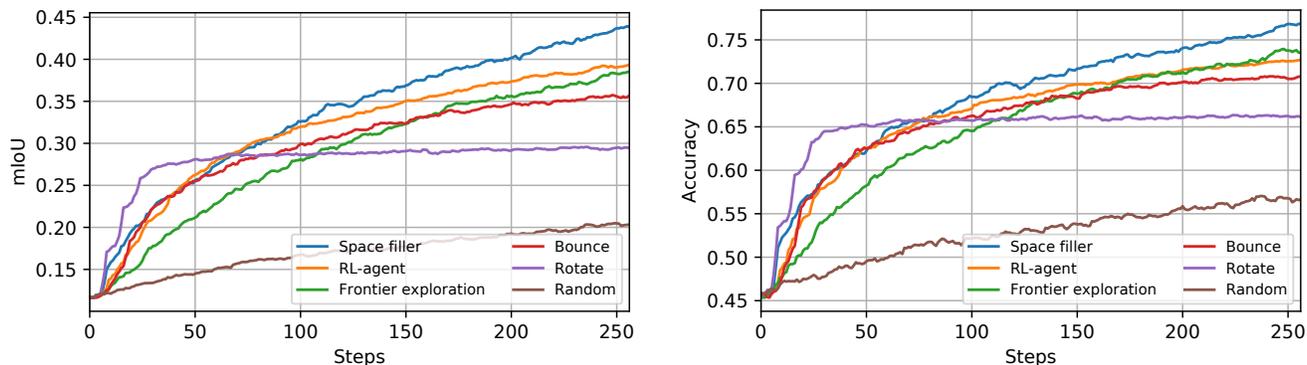


Figure 4: Mean segmentation accuracy and mIoU versus number of actions (steps), evaluated on the Matterport3D test scenes. The RL-agent was trained on 256-step episodes. This agent fairly quickly outperforms all other comparable pre-specified agents. *Rotate* is strong initially since it quickly gathers many annotations in a 360 degree arc, but is eventually outperformed by most other methods that move around in the houses. Frontier exploration yields similar accuracy as the RL-agent after about 170 steps, but uses significantly more annotations (cf. table 1) and assumes perfect pose and depth information. The space filler, which assumes full knowledge of the environment, yields the best results after about 100 steps.

4 Experiments

In this section we provide empirical evaluations of various methods. The primary metrics are mIoU and segmentation accuracy but we emphasize that we test the exploration and annotation selection capability of *policies* – the mIoU and accuracy measure how well agents explore in order to refine their perception. Differently from accuracy, the mIoU does not become overly high by simply segmenting large background regions (such as walls), hence it is more representative of overall semantic segmentation quality.

Experimental setup. We evaluate the methods on the Matterport3D dataset (Chang et al. 2017) using the embodied agent framework Habitat (Savva et al. 2019). This setup allows the agent to freely explore photorealistic 3d models of large houses, that have ground truth annotations for 40 diverse semantic classes. Hence it is a suitable environment for evaluation. To assess the generalization capability of the RL-agent we train and test it in *different* houses. We use the same 61, 11 and 18 houses for training, validation and testing as Chang et al. (2017). The RL-agent and all pre-specified methods except the space filler are *comparable* in terms of assumptions, cf. §3.1. The space filler assumes full spatial knowledge of the environment (ground truth map) and hence has inherent advantages over the other methods.

During RL-agent training we randomly sample starting positions and rotations from the training houses at the start of each episode. An episode ends after 256 actions. Hyperparameters of the learnt and pre-specified agents are tuned on the validation set. For validation and testing we use 3 and 4 starting positions per scene, respectively, so each agent is tested for a total of 33 episodes in validation and 72 episodes in testing. The reported metrics are the mean over all these runs. All methods are evaluated on the same starting positions in the same houses. The reference views used to eval-

uate the semantic segmentation performance are obtained by sampling 32 random views within a 5 m geodesic distance of the agent’s starting position at the beginning of each episode. In training the reference views are sampled randomly. During validation and testing, for fairness, we sample the same views for a given starting position when we test different agents. Note that there is no overlap between reference views during policy training and testing, since training, validation and testing houses are non-overlapping.

Recall that the RL-agent’s policy parameters are denoted by θ . Let θ_{seg} denote the parameters of the underlying semantic segmentation network, in order to clarify when we reset, freeze and refine θ and θ_{seg} , respectively. For RL-training, we refine θ during policy estimation in the training houses. When we evaluate the policy on the validation or test houses we freeze θ and only use the policy for inference. The parameters of the segmentation network θ_{seg} are always reset at the beginning of an episode, regardless of which house we deploy the agent in, and regardless of whether the policy network is training or not. During an episode, we refine θ_{seg} exactly when the agent selects the *Annotate* or *Collect* actions (this applies also to all the other methods described in §3.1). Thus annotated views in an episode are used to refine θ_{seg} in that episode only, and are not used in any other episodes.

4.1 Main Results

We measure the performances of the agents in two settings: (a) with unlimited annotations but limited total actions (max 256, as during RL-training), or (b) for a limited annotation budget (max 100) but unlimited total actions. All methods were tuned on the validation set in a setup similar to (a) with 256 step episodes. Note however that the number of annotations can differ for different methods in a 256 step episode. The setup (b) is used to assess how the different methods compare for a fix number of annotations.

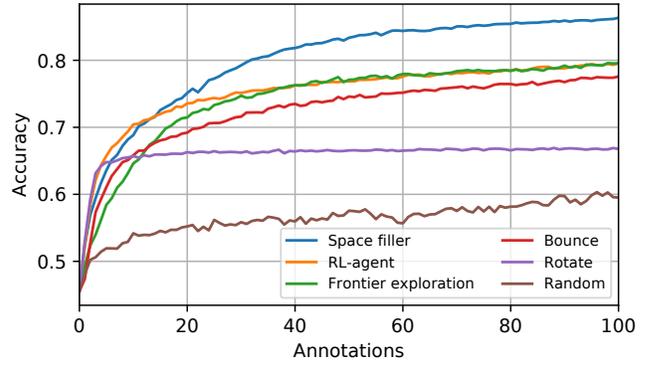
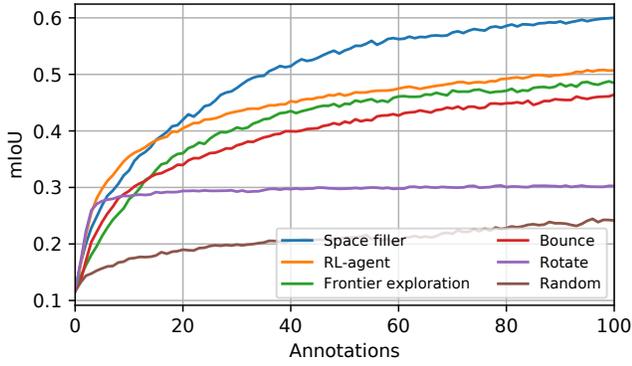


Figure 5: Mean segmentation accuracy and mIoU for a varying number of requested annotations evaluated on the Matterport3D test scenes. The RL-agent outperforms all comparable pre-specified methods (although frontier exploration matches it in accuracy after about 40 annotations), indicating that it has learnt an exploration policy which generalizes to novel scenes. The space filler, as expected, outperforms the RL-agent, except for less than 15 annotations. Thus the RL-agent is best before and around its training regime, where on average annotates 16.7 times per episode, cf. table 1.

Method	mIoU	Acc	# Ann	# Coll
Space filler	0.439	0.769	24.7	23.9
RL-agent	0.394	0.727	16.7	15.2
Frontier exploration	0.385	0.735	24.2	21.6
Bounce	0.357	0.708	29.6	26.0
Rotate	0.295	0.661	34.3	32.7
Random	0.204	0.566	29.1	19.5

Table 1: Comparison of different agents for a fixed episode length of 256 actions on the Matterport3D test scenes. The RL-agent gets higher mIoU using far fewer annotations than comparable pre-specified methods, implying that the RL-agent’s policy selects more informative views to annotate.

Fixed episode length. Table 1 and fig. 4 show results on the test scenes for episodes of length 256. The RL-agent outperforms the comparable pre-specified methods in mIoU and accuracy, although frontier exploration – which uses perfect pose and depth information, and is idealized to always move within the reference view radius – yields similar accuracy after about 170 steps. The RL-agent uses much fewer annotations than other methods, hence those annotated views are more informative. The space filler, which assumes perfect knowledge of the map, outperforms the RL-agent but uses significantly more annotations. Note that the *Rotate* baseline saturates, supporting the intuition that an agent has to move around in order to increase performance in complex environments.

Fixed annotation budget. In table 2 and fig. 5 we show test results when the annotation budget is limited to 100 images per episode. As expected, the space filler yields the best results, although the RL-agent gets comparable performance when using up to 15 annotations. The RL-agent outperforms comparable pre-specified methods in mIoU and accuracy. Frontier exploration obtains similar accuracy.

Method	mIoU	Acc	# Steps	# Coll
Space filler	0.600	0.863	1048	91
RL-agent	0.507	0.796	1541	94
Frontier exploration	0.485	0.796	998	84
Bounce	0.464	0.776	861	87
Rotate	0.303	0.668	752	96
Random	0.242	0.595	910	64

Table 2: Comparison of different agents for a fixed budget of 100 annotations on Matterport3D test scenes. The RL-agent gets a higher mIoU than comparable pre-specified agents, despite not being trained in this setting.

We also see that the episodes of the RL-agent are longer.

Qualitative examples. Fig. 6 shows examples of views that the RL-agent choose to annotate. The agent explores large parts of the space and the annotated views are diverse, both in their spatial locations and in the types of semantic classes they contain. Fig. 7 shows how the segmentation network’s performance on two reference views improves during an episode. The two views are initially poorly segmented, but as the agent explores and acquires annotations for novel views, the accuracy on the reference views increases.

4.2 Ablation Studies of the RL-agent

Ablation results of the RL-agent on the validation set are in table 3. We compare to the following versions: i) Policy without visual features ϕ_{img} ; ii) Policy without ResNet features ϕ_{res} ; iii) No additional exploration reward (4), i.e. $R_t^{exp} = 0$; iv) No *Collect* action and \mathbf{P}_t is not an input to ϕ_{img} ; and v) Only exploration trained, using the heuristic strategy for annotations. We trained the ablated models for 4,000 episodes as for the full model.

Both the validation accuracy and mIoU are higher for the full RL-model compared to all ablated variants, justifying



Figure 6: The first six requested annotations by the RL-agent in a room from the test set. Left: Map showing the agent’s trajectory and the six first requested annotations (green arrows). The initially given annotation is not indicated with a number. Blue arrows indicate `Collect` actions. Right: For each annotation (numbered 1 - 6) the figures show the image seen by the agent and the ground truth received when the agent requested annotations. As can be seen, the agent quickly explores the room and requests annotations containing diverse semantic classes.

Variant	mIoU	Acc	# Ann	# Coll
Full model	0.427	0.732	16.4	16.4
No collect nor P_t	0.415	0.727	17.9	0.0
Only exploration	0.411	0.727	16.1	14.4
$R_t^{exp} = 0$	0.401	0.719	17.7	47.4
No ϕ_{img}	0.378	0.696	14.3	3.8
No ResNet	0.375	0.705	23.3	0.3

Table 3: Ablation study of different RL-based model variants for 256-step episodes on the validation set. The full RL-agent outperforms all ablated models at a comparable or lower number of requested annotations.

design choices. The model not relying on propagating annotations and using the `Collect` action performs somewhat worse than the full model despite a comparable amount of annotations. The learnt annotation strategy yields higher mIoU and accuracy compared to the heuristic one, at comparable number of annotations. The exploration reward is important in encouraging the agent to navigate to unvisited positions – without it performance is worse, despite a comparable number of annotations. The agent trained without the exploration reward uses an excessive number of `Collect` actions, so this agent often stands still instead of moving. Finally, omitting either visual or ResNet features from the policy significantly harms accuracy for the resulting recognition system.

4.3 Analysis of Annotation Strategies

In this section we examine how different annotation strategies affect the task performance on the validation set for the space filler and bounce methods. Specifically, the annotation strategies are:

- **Threshold perception.** This is the variant evaluated in §4.1, i.e. it issues the `Collect` action when 30% of the propagated labels are unknown and `Annotate` when 85% are unknown.
- **Learnt perception.** We train a simplified RL-agent where

Variant	mIoU	Acc	# Ann	# Coll
Threshold perception	0.472	0.770	20.8	19.9
Learnt perception	0.454	0.755	22.8	37.4
Random perception	0.446	0.747	24.2	24.4

Table 4: Results for different model variants of the space filler method. We report the mean on the validation scenes. The threshold perception strategy – which is the one used in the main evaluations in §4.1 – yields the best results.

the movement actions are restricted to follow the exploration trajectory of the baseline method (space filler and bounce, respectively). This model has 3 actions: move along the baseline exploration path, `Annotate` and `Collect`. All other training settings are identical to the full RL-agent.

- **Random perception.** In each step, this variant follows the baseline exploration trajectory with 80% probability, while annotating views and collecting propagated labels with 10% probability each.

As can be seen in table 4, the best results for the space filler are obtained by using the threshold strategy, which also annotates slightly less frequently than other variants. Using learnt perception actions yields better results compared to random perception actions, and takes slightly fewer annotations per episode. Similar results carry over to the bounce method in table 5, i.e. the best results are again obtained by the threshold variant. The model with a learnt annotation strategy fails to converge to anything better than heuristic perception strategies. In fact, it converges to selecting `Collect` almost 40% of the time, which indicates a lack of movement for this variant.

In table 3 we saw that a learnt exploration method with a heuristic annotation strategy yields worse results than a fully learnt model. Conversely, the results from table 4 and table 5 show that a heuristic exploration method using a learnt annotation strategy yields worse results than an entirely heuristic model. Together these results indicate that it is necessary to

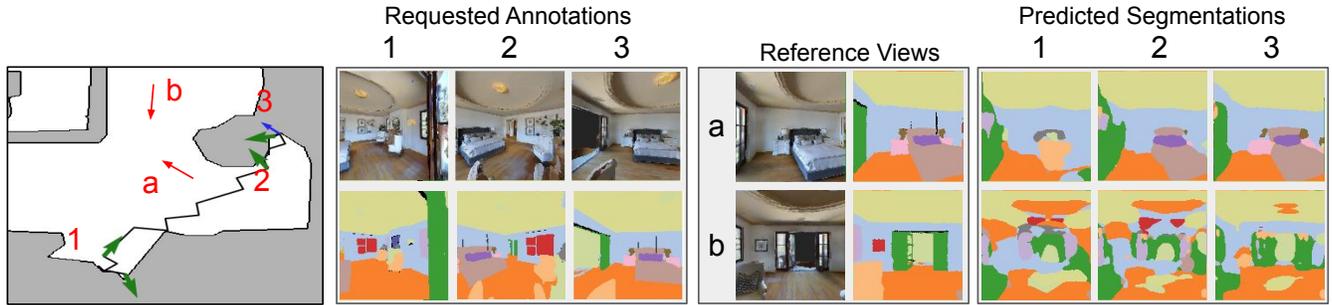


Figure 7: Example of the RL-agent’s viewpoint selection and how its perception improves over time. We show results of two reference views after the first three annotations of the RL-agent. Left: Agent’s movement path is drawn in black on the map. The annotations (green arrows) are numbered 1 - 3, and the associated views are shown immediately right of the map (the initially given annotation is not shown). Red arrows labeled *a* - *b* indicate the reference views. Right: Reference views and ground truth masks, followed by predicted segmentation after one, two and three annotations. Notice clear segmentation improvements as the agent requests more annotations. Specifically, note how reference view *a* improves drastically with annotation 2 as the bed is visible in that view, and with annotation 3 where the drawer is seen. Also note how segmentation improves for reference view *b* after the door is seen in annotation 3.

Variant	mIoU	Acc	# Ann	# Coll
Threshold perception	0.388	0.706	27.4	24.5
Learnt perception	0.375	0.699	14.6	98.8
Random perception	0.379	0.698	25.9	24.6

Table 5: Results for different model variants of the bounce method. We report the mean on the validation scenes. The threshold perception strategy – which is the one used in the main evaluations in §4.1 – yields the best results, but also uses the largest amount of annotations on average.

learn how to annotate and explore jointly to provide the best results, given comparable environment knowledge.

4.4 Pre-training the Segmentation Network

Recall that our semantic segmentation network is randomly initialized at the beginning of each episode. In this section we evaluate the effect of instead pre-training the segmentation network² on the 61 training houses using about 20,000 random views. In table 6 we compare using this pre-trained segmentation network as initialization for the RL-agent with the case of random initialization. We also show results when not further fine-tuning the pre-trained segmentation network, i.e. when not performing any embodied visual active learning.

The weak result obtained when not fine-tuning (first row) indicates significant appearance differences between the houses. This is further suggested by the fact that the RL-agent gets a surprisingly modest boost from pre-training the segmentation network (third row vs second row). Note the different number of annotated views used here – the agent without pre-training uses only 16.4 views on average, while the other uses about 20,000 + 14.4 annotated views, if we

²In this pre-training experiment, we use the same architecture and hyperparameters for the segmentation network as when it is trained and deployed in the embodied visual active learning task.

Variant	mIoU	Acc	# Ann	# Coll
Pre-train, no RL	0.208	0.549	20k	0.0
No pre-train, RL	0.427	0.732	16.4	16.4
Pre-train, RL	0.461	0.780	20k + 14.4	13.3

Table 6: Results for different training regimes for the semantic segmentation network. A pre-trained segmentation network generalizes poorly to unseen environments (first row), and there is relatively little gain for the RL-agent by having a pre-trained segmentation network (third row). Note that pre-training uses over 1000x more annotations compared to performing embodied active visual learning from scratch.

count all the images used for pre-training. Due to relatively marginal gains for a large number of annotated images, we decided to evaluate all agents without pre-training the segmentation network.

5 Conclusions

In this paper we have explored the *embodied visual active learning* task for semantic segmentation and developed a diverse set of methods, both pre-designed and learning-based, in order to address it. The agents can explore a 3d environment and improve the accuracy of their semantic segmentation networks by requesting annotations for informative viewpoints, propagating annotations via optical flow at no additional cost by moving in the neighborhood of those views, and self-training. We have introduced multiple baselines as well as a more sophisticated fully learnt model, each exposing different assumptions and knowledge of the environment. Through extensive experiments in the photorealistic Matterport3D environment we have thoroughly investigated the various methods and shown that the fully learning-based method outperforms comparable non-learnt approaches, both in terms of accuracy and mIoU, while relying on fewer annotations.

Acknowledgments

This work was supported in part by the European Research Council Consolidator grant SEED, CNCS-UEFISCDI PN-III-P4-ID-PCE-2016-0535 and PCCF-2016-0180, the EU Horizon 2020 Grant DE-ENIGMA, Swedish Foundation for Strategic Research (SSF) Smart Systems Program, as well as the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

- Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.
- Ammirato, P.; Poirson, P.; Park, E.; Košecák, J.; and Berg, A. C. 2017. A dataset for developing and benchmarking active vision. In *ICRA*, 1378–1385. IEEE.
- Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.
- Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision*.
- Chaplot, D. S.; Gandhi, D.; Gupta, S.; Gupta, A.; and Salakhutdinov, R. 2020. Learning To Explore Using Active Neural SLAM. In *ICLR*.
- Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI* 40(4): 834–848.
- Chen, T.; Gupta, S.; and Gupta, A. 2019. Learning exploration policies for navigation. In *ICLR*.
- Das, A.; Datta, S.; Gkioxari, G.; Lee, S.; Parikh, D.; and Batra, D. 2018. Embodied question answering. In *CVPR*, volume 5, 6.
- Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An Open Urban Driving Simulator. In *CoRL*, 1–16.
- Fang, K.; Toshev, A.; Fei-Fei, L.; and Savarese, S. 2019. Scene Memory Transformer for Embodied Agents in Long-Horizon Tasks. In *CVPR*.
- Fang, M.; Li, Y.; and Cohn, T. 2017. Learning how to Active Learn: A Deep Reinforcement Learning Approach. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 595–605.
- Feng, D.; Wei, X.; Rosenbaum, L.; Maki, A.; and Dietmayer, K. 2019. Deep Active Learning for Efficient Training of a LiDAR 3D Object Detector. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, 667–674.
- Gärtner, E.; Pirinen, A.; and Sminchisescu, C. 2020. Deep Reinforcement Learning for Active Human Pose Estimation. In *AAAI*, 10835–10844.
- Gupta, S.; Davidson, J.; Levine, S.; Sukthankar, R.; and Malik, J. 2017. Cognitive mapping and planning for visual navigation. In *CVPR*, 2616–2625.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.
- Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.
- Jayaraman, D.; and Grauman, K. 2016. Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion. In *ECCV*, 489–505. Springer.
- Jayaraman, D.; and Grauman, K. 2018. Learning to look around: Intelligently exploring unseen environments for unknown tasks. In *CVPR*.
- Jia Deng; Wei Dong; Socher, R.; Li-Jia Li; Kai Li; and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. ISBN 978-1-4244-3992-8. doi:10.1109/CVPRW.2009.5206848.
- Johns, E.; Leutenegger, S.; and Davison, A. J. 2016. Pairwise decomposition of image sequences for active multi-view recognition. In *CVPR*, 3813–3822.
- Kingma, D.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kolve, E.; Mottaghi, R.; Han, W.; VanderBilt, E.; Weihs, L.; Herrasti, A.; Gordon, D.; Zhu, Y.; Gupta, A.; and Farhadi, A. 2017. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv preprint arXiv:1712.05474*.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. In *NeurIPS*.
- Liang, E.; Liaw, R.; Nishihara, R.; Moritz, P.; Fox, R.; Goldberg, K.; Gonzalez, J.; Jordan, M.; and Stoica, I. 2018. RLlib: Abstractions for distributed reinforcement learning. In *International Conference on Machine Learning*, 3053–3062.
- Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.
- Lughofer, E. 2012. Single-pass active learning with conflict and ignorance. *Evolving Systems* 3(4): 251–271.
- Mishkin, D.; Dosovitskiy, A.; and Koltun, V. 2019. Benchmarking Classic and Learned Navigation in Complex 3D Environments. *arXiv preprint arXiv:1901.10915*.
- Mousavian, A.; Toshev, A.; Fišer, M.; Košecák, J.; Wahid, A.; and Davidson, J. 2019. Visual representations for semantic target driven navigation. In *ICRA*, 8846–8852. IEEE.
- Pardo, A.; Xu, M.; Thabet, A.; Arbelaez, P.; and Ghanem, B. 2019. BAOD: Budget-Aware Object Detection. *arXiv preprint arXiv:1904.05443*.
- Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven exploration by self-supervised prediction. In *ICML*, volume 2017.
- Pot, E.; Toshev, A.; and Kosecka, J. 2018. Self-supervisory Signals for Object Discovery and Detection. *arXiv preprint arXiv:1806.03370*.

- Qi, W.; Mullapudi, R. T.; Gupta, S.; and Ramanan, D. 2020. Learning to Move with Affordance Maps. In *ICLR*.
- Ramakrishnan, S. K.; Jayaraman, D.; and Grauman, K. 2020. An Exploration of Embodied Visual Exploration. *arXiv preprint arXiv:2001.02192*.
- Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *CVPR*, 779–788.
- Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 91–99.
- Savva, M.; Chang, A. X.; Dosovitskiy, A.; Funkhouser, T.; and Koltun, V. 2017. MINOS: Multimodal Indoor Simulator for Navigation in Complex Environments. *arXiv:1712.03931*.
- Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; et al. 2019. Habitat: A platform for embodied ai research. In *ICCV*, 9339–9347.
- Sax, A.; Emi, B.; Zamir, A. R.; Guibas, L. J.; Savarese, S.; and Malik, J. 2019. Mid-Level Visual Representations Improve Generalization and Sample Efficiency for Learning Visuomotor Policies. In *CoRL*.
- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Settles, B. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Song, S.; Zeng, A.; Chang, A. X.; Savva, M.; Savarese, S.; and Funkhouser, T. 2018. Im2pano3d: Extrapolating 360 structure and semantics beyond the field of view. In *CVPR*, 3847–3856.
- Song, S.; Zhang, L.; and Xiao, J. 2015. Robot in a room: Toward perfect object recognition in closed environments. *CoRR*, abs/1507.02703.
- Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. Pwcnet: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 8934–8943.
- Sundaram, N.; Brox, T.; and Keutzer, K. 2010. Dense point trajectories by GPU-accelerated large displacement optical flow. In *ECCV*, 438–451. Springer.
- Wang, K.; Lin, Y.; Wang, L.; Han, L.; Hua, M.; Wang, X.; Lian, S.; and Huang, B. 2019. A unified framework for mutual improvement of slam and semantic segmentation. In *ICRA*, 5224–5230. IEEE.
- Wijmans, E.; Datta, S.; Maksymets, O.; Das, A.; Gkioxari, G.; Lee, S.; Essa, I.; Parikh, D.; and Batra, D. 2019. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. In *CVPR*.
- Woodward, M.; and Finn, C. 2017. Active one-shot learning. *arXiv preprint arXiv:1702.06559*.
- Xia, F.; R. Zamir, A.; He, Z.-Y.; Sax, A.; Malik, J.; and Savarese, S. 2018. Gibson env: real-world perception for embodied agents. In *CVPR*. IEEE.
- Xiong, B.; and Grauman, K. 2018. Snap angle prediction for 360 panoramas. In *ECCV*, 3–18.
- Yamauchi, B. 1997. A frontier-based approach for autonomous exploration. In *CIRA*, 146–151. IEEE.
- Yang, H.-K.; Chiang, P.-H.; Ho, K.-W.; Hong, M.-F.; and Lee, C.-Y. 2019a. Never Forget: Balancing Exploration and Exploitation via Learning Optical Flow. *arXiv preprint arXiv:1901.08486*.
- Yang, J.; Ren, Z.; Xu, M.; Chen, X.; Crandall, D. J.; Parikh, D.; and Batra, D. 2019b. Embodied Amodal Recognition: Learning to Move to Perceive Objects. In *ICCV*, 2040–2050.
- Yu, L.; Chen, X.; Gkioxari, G.; Bansal, M.; Berg, T. L.; and Batra, D. 2019. Multi-Target Embodied Question Answering. In *CVPR*.
- Zheng, L.; Zhu, C.; Zhang, J.; Zhao, H.; Huang, H.; Niessner, M.; and Xu, K. 2019. Active scene understanding via online semantic reconstruction. In *Computer Graphics Forum*, volume 38, 103–114. Wiley Online Library.
- Zhong, F.; Wang, S.; Zhang, Z.; and Wang, Y. 2018. Detectslam: Making object detection and slam mutually beneficial. In *WACV*, 1001–1010. IEEE.
- Zhu, Y.; Mottaghi, R.; Kolve, E.; Lim, J. J.; Gupta, A.; Fei-Fei, L.; and Farhadi, A. 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*, 3357–3364. IEEE.