IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.1120000

Interactive Semantic Map Representation for Skill-Based Visual Object Navigation

TATIANA ZEMSKOVA[2], ALEKSEI STAROVEROV[1], KIRILL MURAVYEV[3], DMITRY YUDIN[1,2],ALEKSANDR PANOV[1,2,3]

[1]Artificial Intelligence Research Institute (AIRI), 32 Kutuzovsky Ave., Moscow, 121170, Russia
[2]Moscow Institute of Physics and Technology, 9 Institutsky per., Dolgoprudny, 141701, Russia
[3]Federal Research Center "Computer Science and Control", 9 60-Letiya Oktyabrya Ave., Moscow, 117312, Russia

Corresponding author: Tatiana Zemskova (e-mail: zemskova.ts@phystech.edu).

This work was supported by Russian Science Foundation, grant No. 20-71-10116, https://rscf.ru/en/project/20-71-10116/.

ABSTRACT Visual object navigation is one of the key tasks in mobile robotics. One of the most important components of this task is the accurate semantic representation of the scene, which is needed to determine and reach a goal object. This paper introduces a new representation of a scene semantic map formed during the embodied agent interaction with the indoor environment. It is based on a neural network method that adjusts the weights of the segmentation model with backpropagation of the predicted fusion loss values during inference on a regular (backward) or delayed (forward) image sequence. We implement this representation into a full-fledged navigation approach called SkillTron. The method can select robot skills from end-to-end policies based on reinforcement learning and classic map-based planning methods. The proposed approach makes it possible to form both intermediate goals for robot exploration and the final goal for object navigation. We conduct intensive experiments with the proposed approach in the Habitat environment, demonstrating its significant superiority over state-of-the-art approaches in terms of navigation quality metrics. The developed code and custom datasets are publicly available at github.com/AIRI-Institute/skill-fusion.

INDEX TERMS semantic map; navigation; robotics; reinforcement learning; frontier-based exploration

I. INTRODUCTION

Accurate visual navigation to target objects in unfamiliar environments is crucial for on-board mobile robot systems. In this case, the selection of embodied agent actions can be performed by various methods: classical modular map-based motion planning algorithms [1], or end-to-end neural network models based on images from on-board cameras and/or image segmentation masks [2]. Classical approaches use separate modules to build a map, explore the environment, select the final goal, and plan a path to the goal. End-to-end models are typically trained using reinforcement learning methods, which are valuable whenever information about the environment is incomplete.

All visual navigation methods use a semantic representation of the surrounding scene, typically formed using various trained image segmentation models.

A multi-channel segmentation mask can be used directly as the representation of a one-shot observed map. Another approach involves building an accumulated map representation in the form of bird's-eye-view 2D [3], 2.5D [4] or 3D [5] semantic maps, where each map point contains a class label of the found object.

The possibility of increasing the efficiency of visual navigation methods based on semantic maps is studied in special photorealistic simulators. For indoor navigation, some of the most popular environments are Habitat [6], AI2-THOR [7], OmniGibson [8]. For the study of outdoor navigation, there are environments such as Carla [9], AirSim [10], etc. Isaac Sim [11] is one example of a universal simulator that can be used both indoors and outdoors. This paper considers only simulation environments for indoor navigation to given objects.

Contemporary research [3], [12], as well as solutions of such indoor navigation competitions as the Habitat Challenge[1], shows that the semantic representation of the map plays a key role in increasing the navigation efficiency. Typical approaches map semantic segmentation results using heuristic projection algorithms and noisy information about the agent's position [13]. Some other approaches [14] involve straightforward multi-class semantic mapping and heuristic

[1] https://aihabitat.org

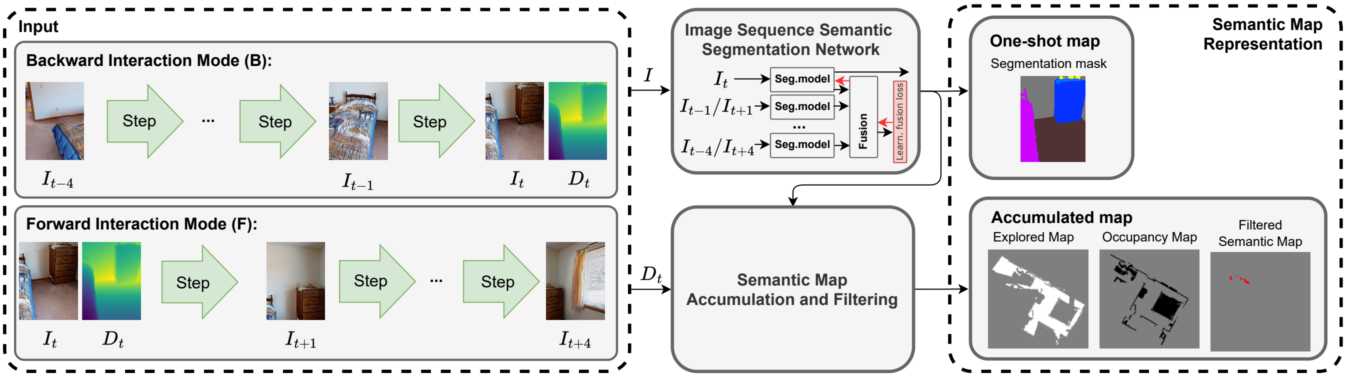


FIGURE 1. **Proposed two-level representation of a semantic map. The one-shot map in the form of a segmentation mask is built during the interaction of the embodied agent with the environment. We consider regular (backward) or delayed (forward) image sequences as a result of such interaction. The accumulated map is a multi-channel semantic bird's-eye-view map formed using an original filtering algorithm.**

path planners. However, the results on the success of reaching target objects remain rather scarce. For example, the success rate of the state-of-the-art methods for object navigation does not exceed 60% in the Habitat Challenge 2022 and 2023 competitions[2]. Therefore, new integrated approaches to the semantic map representation are needed. These approaches should take into account the specifics of the moving embodied agent.

The quality of semantic map representation directly depends on the segmentation quality of 2D observation from the camera at each time point. The agent's field of view is limited, so the semantic segmentation may contain errors. For example, armchairs and sofas cannot be distinguished from each other from certain angles. Such semantic errors may be corrected as the agent moves through space and the object is being observed from different view angles. To solve this problem, the recent methods for semantic mapping [15], [16], compute the multi-class probability distribution for each point in a semantic map. The probability distribution can be interactively updated over time based on new observations [15]. In the case of a large vocabulary of semantic classes, i.e. over 100 classes, such a map representation becomes extensively voluminous.

Without using any information compression methods, $O(NK_s)$ memory is required to store information about the distribution of classes in the semantic map representation of the environment. Here N is the number of elements (e.g. points, voxels, grid cells, and graph nodes) in the semantic map representation, and K_s is the class number of the distribution stored in memory. Accordingly, adding a new class will increase the size of memory consumption by $O(N)$.

Therefore, we use a different approach to refine the semantic map during navigation. We fuse the information from an image sequence into a single multi-channel semantic segmentation mask using a neural network model. An image sequence provides more contextual information about the navigation scene. In this case, increasing the class number primarily affects the amount of memory required to represent the 2D one-shot semantic map in the form of the segmentation mask. The accumulated map of the navigation environment is stored in a compact form of a 2D bird's-eye-view semantic map, which contains information only about the target class and the classes related to it. Notably, previous methods [15], [17] use 3D semantic map representations.

We propose a full-fledged navigation approach called SkillTron that uses adjustment of the segmentation model weights with backpropagation of the predicted fusion loss values during inference on a regular (backward) or delayed (forward) image sequence. SkillTron can select robot skills from end-to-end policies based on reinforcement learning and classic map-based planning methods.

Our main contributions are:

- We propose a two-level representation of a semantic map (refer to Figure 1) that aims to reduce semantic noise in two stages. A one-shot map in the form of a segmentation mask is constructed during the interaction of the embodied agent with the environment. At this stage, additional observations are used to refine the 2D segmentation mask. The accumulated map is formed using a unique filtering algorithm for the semantic bird's-eye-view map.
- We develop a navigation approach, called SkillTron, using the proposed interactive semantic map representation with robot skill selection from end-to-end policies based on reinforcement learning and traditional map-based planning methods. This allows us to form both intermediate goals for robot exploration and the final goal for object navigation.
- To evaluate the proposed approach, we gather several custom datasets in the photo-realistic Habitat Indoor Environment. We utilize these datasets to train and test the interactive semantic segmentation model as well as to directly assess the quality of the approach for building the accumulated map. During the experiments, we demonstrate that our approach significantly improves the quality of semantic map representation. The proposed SkillTRon approach outperforms current state-of-

[2] https://aihabitat.org/challenge/2023/

the-art methods for the indoor Object Goal task.

II. RELATED WORKS

A. NAVIGATION

Like most navigational tasks, the ObjectNav task can be addressed using Simultaneous Localization and Mapping (SLAM) techniques and deterministic path planners. As a result, the agent constructs an occupancy map and a collision-free path to the goal. Because of the unknown coordinates of the goal object, methods like Frontier-based exploration (FBE) [18] are frequently employed. A frontier is the boundary between the explored free and unexplored spaces. A point on the frontier is selected as the goal, and the robot navigates to it to explore new space. Most frontier-based exploration algorithms, like [1], choose an exploration goal point on the frontiers according to a complex heuristic cost function. If the agent sees a goal type of object during this exploration, it navigates directly to it.

An alternative approach to exploration is predicting a potential function for each map cell to choose a goal. Such approach is used in the PONI method [14] whereby a two-component potential function is predicted by a UNet-like neural network directly from a multi-layer 2D bird's-eye-view map. The map layers include an obstacle layer, an explored space layer, and 16 layers for semantic classes. The first component of a point's potential is the estimation of an area which can be explored from the point; the second component is the probability of a goal object present near the point. The potential function approach enables goal-oriented exploration; also, differences in image parameters do not affect the prediction quality because the neural network takes a 2D map as input. However, like any map-based approach, it has some drawbacks: it is affected by the mapping noise and by the semantic segmentation noise. So, the agent may stuck in an unmapped obstacle or navigate to a spuriously mapped goal object.

Another large family of methods use end-to-end learning-based approaches, mostly Reinforcement Learning (RL) techniques. Learnable methods are able to quickly explore spaces without hitting obstacles because they directly use information from images and learn directly for solving exploration task effectively. A significant breakthrough of the learning approaches in navigation tasks was the DDPPO method [19], which used Proximal Policy Optimization [20] at its core. Without mapping or planning modules, DDPPO solves the PointNav task at a level of human performance after training on 2.5 billion steps in the environment.

However, DDPPO demonstrates a poor performance at the ObjectNav task. The pure end-to-end RL algorithms with vanilla visual and recurrent modules perform insufficiently due to overfitting and sample inefficiency. The authors of the Auxiliary task RL method [21] partially solve this problem by adding auxiliary learning tasks and an exploration reward during the training phase. However, even with such enhancement, RL methods ‘‘forget’’ scene information after a certain number of steps due to limitations in the recurrent network layers. So, the agent starts re-exploring previously visited areas, reducing navigation efficiency.

A promising approach to solving the ObjectNav task is to mix analytic and learned components and operate on explicit spatial maps of the environment. Such a combination of classical and learned methods was employed in the SemExp [13], CoW [12], and SkillFusion [3] methods. Typically, authors use a deterministic map module and divide a policy into a global one that outputs a short-term subgoal by planning on a map and a local one that pursues that subgoal. Drawing upon those works, we also build a two-level policy. However, our low-level policy consists of several independent skills and high-level policy switches between them depending on what the agent needs to do at a given time.

B. INTERACTIVE COMPUTER VISION

An embodied agent navigating through the environment can interactively update the scene's semantic representation based on new sensor data. This allows the agent to improve the semantic understanding of the scene. The recent emergence of environments for embodied agents, e.g. Habitat [6], AI2-THOR [7], and OmniGibson [8], enabling the simulation of agent navigation and interaction with different objects, led to the development of interactive segmentation methods.

An agent can predict its future actions to improve perception quality for the next observation. In this case, interactive segmentation would be a special case of the Next Best View selection task aiming to identify the next most informative sensor position for computer vision tasks. Recent papers propose different methods to assess the informative value of a view. The choice of the next best view can rely on the confidence score of a frozen object detector [22], segmentation quality [23], [24], or statistical criteria derived from the image itself [25]. The interactive computer vision methods use both learned policies [22]–[24] and predefined policies, e.g. the output of a voting system [25], for the next best view selection.

Other interactive segmentation methods aim to improve the quality of the semantic representation of an environment rather than focusing solely on object recognition. The VLN-SIG method [26] improves the agent's performance on the Vision-Language Navigation Task by adding an auxiliary task to predict the view semantics for the next step. The authors of the SSMI method [15] propose learning an exploration policy to decrease the uncertainty of different semantic classes by considering the motion cost.

We can highlight the methods that improve the understanding of the semantics of a scene based on active exploration. The embodied agents use active exploration to facilitate the adaptation to complex and unfamiliar environments. Agents can query human expert help [27] or create pseudo-labels in testing environments using multiple points of view [28]. A different approach involves introducing learnable policies that consider the uncertainty of semantic maps for collecting data to fine-tune semantic segmentation models of embodied agents [17], [29] .

Another approach for adapting models to a test environment is fine-tuning the model during inference. Recent works Interactron [30] and SegmATRon [31] propose the use of an adaptive loss function predicted from a frame sequence to refine object detection [30] and semantic segmentation [31]. The adaptive loss function is used during both training and inference. Our work modifies the SegmATRon approach to create an interactive semantic map representation. We investigate different methods for selecting consecutive RGB observations to improve the quality of the semantic map representation.

III. OBJECT GOAL NAVIGATION TASK

In the context of the indoor Object Goal task as described in the literature [32], the aim is to navigate toward an instance of a specified object category $C \in \{c_1, c_2, ..., c_n\}$ (e.g., a *chair*) within an unfamiliar environment. The agent receives an observation $S = (S_{RGBD}, S_{GPS+Compass}, C)$ at each step. The action space is discrete and encompasses four types of actions:

- `callstop` to terminate the episode,
- `forward` by 0.25m,
- `turnleft`,
- `turnright` by angle α.

In our experiments, the turn angle α can be equal to 30° or 15°.

The choice of such discrete actions is typical of indoor simulators such as Habitat [6].

After the agent executes the `callstop` action, the agent's performance is assessed using three primary metrics:

1) Success, where an episode is deemed successful if the agent executes the `callstop` command within 1.0m of any object of the goal type;
2) Success weighted, i.e. inverse normalized, by Path Length (SPL), where success is weighted by the efficiency of the agent's path with respect to the shortest path to the nearest goal object from the starting point;
3) SoftSPL, where binary success is substituted by progress toward the goal.

IV. METHOD

A. SKILLTRON NAVIGATION APPROACH

The Object Goal Navigation task poses a significant challenge due to the visual diversity of the scenes. When placed in a new scene without any prior information, an agent struggles to locate the goal.

This challenge is amplified when employing an end-to-end Reinforcement Learning (RL) agent, given the intricacies of the reward model. We test two versions of reward functions. The first, a sparse reward function, provides the agent with a positive reward only upon reaching the goal. The second, a dense reward function, rewards the agent in proportion to the reduction in distance to the nearest goal.

The first version is not sample-efficient due to the problem of getting first positive results. The second version has a problem of penalizing not reaching the closest goal that leads to unstable behavior.

Our approach (see Fig. 2) involves bifurcating the ObjectNav task into two discrete skills: the Exploration skill and the GoalReacher skill. The Exploration skill is designed to locate the goal and represent it on a map. Conversely, the GoalReacher skill is intended to navigate toward a goal within a specified distance once the agent has visual confirmation of the goal from afar. To avoid segmentation outliers and eliminate the noise, we implement semantic map filtering with erosion, dilation, accumulation, and fading.

B. MAP-BASED SKILLS.

In our research, we utilize the Potential Function for ObjectGoal Navigation (PONI [14]) method as a map-based Exploration skill. The PONI method navigates the environment using a multi-layered map, which comprises the explored space layer, the obstacle layer, and the layers for 16 semantic classes. This map is constructed during the exploration process using depth observations and semantic segmentation masks.

For exploration purposes, PONI establishes intermediate goals using a learning-based potential function on the multi-layered map. This potential function is composed of two components: the area potential and the object potential. The area potential indicates the extent of the area that can be explored from a point on the map, while the object potential represents the proximity to the goal object. The potential function neural network is based on an UNet-like map encoder [33], and includes two UNet decoders for the area and object components.

In the context of autonomous navigation, when a goal object is projected onto the semantic map, a path to the nearest cell of the object is constructed utilizing the A-Star path planning algorithm [3], [34]. The robot then navigates along this planned path using a straightforward path follower. The planning algorithm and the path follower constitute the classical GoalReacher skill. The navigation episode is deemed complete when the robot's proximity to the target entity falls below a predetermined threshold. In the context of our experimental framework, this threshold was established at 0.84.

C. LEARNING-BASED SKILLS.

As regards the learning-based skills, we employ a learning-based approach within the framework of a Markov Decision Process (MDP), formally represented as $< S, A, T, R, \gamma >$. Here, S represents the set of states, A signifies the set of possible actions, $T(s_{t+1}|s_t, a_t)$ is the transition function, R is the reward function, and γ is the discount factor.

To distinguish between the Exploration and GoalReacher skills, we have defined separate reward functions for each. The Exploration skill receives a reward of $+1$ for each newly visited $1m^2$ area. The GoalReacher skill, on the other hand, receives a reward relative to the reduction in distance to the observed object.

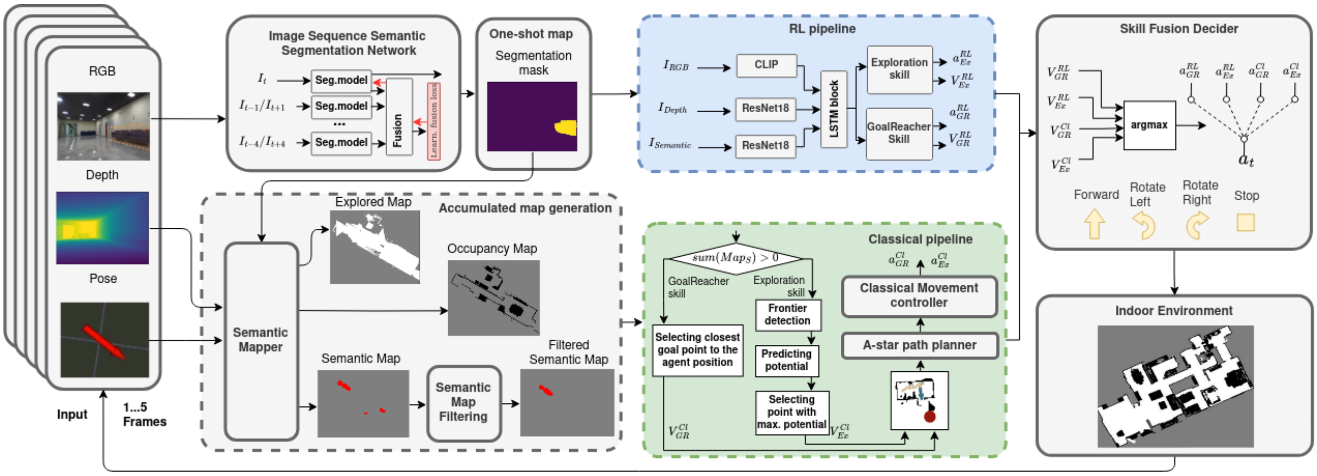


FIGURE 2. A scheme of the proposed SkillTRon visual navigation approach. SkillTron generates robot actions during interaction with the indoor environment. We use a fusion of robot skills, which are formed using a classic modular pipeline and end-to-end RL-based policies. Both of these pipelines utilize different layers of semantic map representation: the first uses an accumulated multi-layer map, whereas the second works with a one-shot environment map.

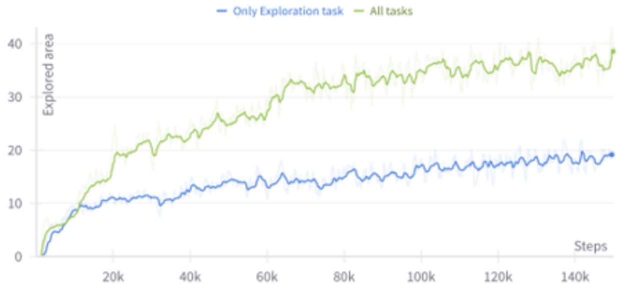


FIGURE 3. The comparison of the area explored using the reinforcement learning (RL) exploration skill when training on a single task versus multiple tasks.

Instead of training two separate algorithms, we have integrated them into a single neural network with two heads, using a late fusion approach. This allows us to develop a dynamic and robust RNN encoder, as each skill depends on different parts of the observation, all of which are retained in the joint LSTM block (see Fig. 3).

For the training of these learning-based skills, we employ Proximal Policy Optimization (PPO) [20] with Generalized Advantage Estimation. We have set the discount factor γ to 0.99 and the GAE parameter τ to 0.95. Each worker collects up to 128 steps of experience from 24 agents running in parallel. Twelve out of these 24 agents use the Exploration skill reward function. The remaining 12 agents use the GoalReacher skill reward function. After the experience collection, each worker performs four epochs of PPO with 2 mini-batches per epoch. We use the Adam optimization algorithm with a learning rate of 5.0×10^{-4}, without normalizing advantages.

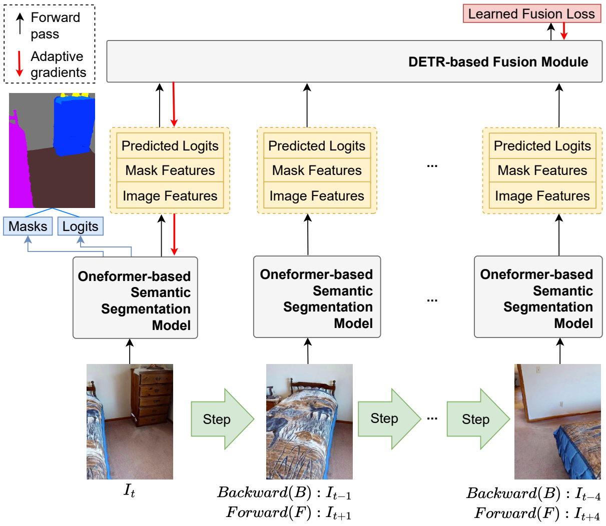


FIGURE 4. Image Sequence Semantic Segmentation Network: SegmATRon (B) and SegmATRon (F).

D. SKILL FUSION DECIDER

At each step, both the map-based and RL approaches are updated based on the observations. The navigational actions are determined by one of these approaches, depending on the value functions of the skills. During the exploration stage, our skill fusion decider alternates between the map-based and RL-based approaches by comparing the RL critic score V^{RL}_{Ex} with a PONI potential function V^{Cl}_{Ex}. During the goal-reaching stage, we assign the classical GoalReacher skill the maximum possible constant value V^{Cl}_{GR} if the goal is mapped and the classical planner can construct a path to it. If this is not the case, we utilize the RL-based GoalReacher critic head to estimate the GoalReacher skill value V^{RL}_{GR}.

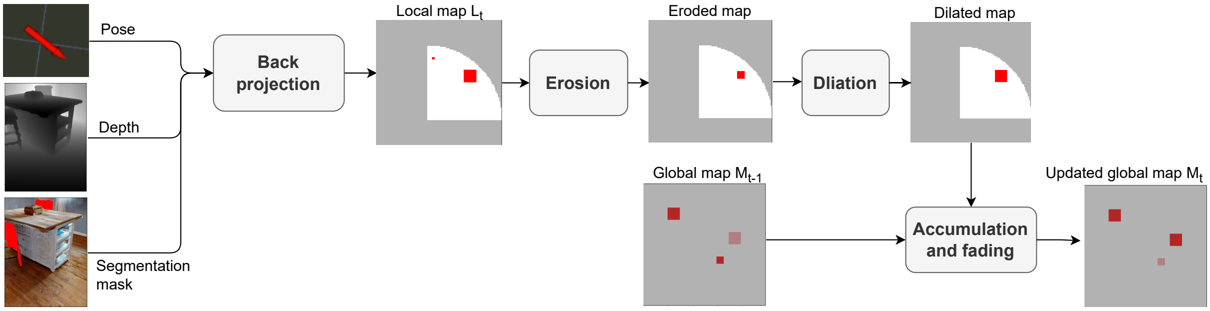


FIGURE 5. Scheme of the proposed semantic map filtering.

E. INTERACTIVE SEGMENTATION

As a baseline method for interactive image segmentation, we consider the SegmATRon model [31] fusing the information from several frames through the mechanism of a hybrid multicomponent fusion loss function. The detailed illustration of the SegmATRon is presented in Fig. 4. The SegmATron model consists of two modules: a semantic segmentation model and a fusion module. We choose the OneFormer [35] modification as the semantic segmentation model. The fusion module is represented by the DETR Transformer Decoder [36]. For each frame of the input image sequence, the semantic segmentation model outputs image features, mask features, and predicted logits. The sequence of outputs is passed to the Fusion Module. The Fusion module predicts the learned fusion loss $\mathcal{L}_{fusion}$ that is used to update the parameters of OneFormer. Then, the updated OneFormer makes another prediction for the first frame in the sequence. The predicted masks and logits are considered as final semantic segmentation of the first frame in the sequence. Fig. 4 illustrates the SegmATRon [31] inference process on an image sequence.

The adaptive fusion loss function $\mathcal{L}_{fusion}(\phi, \theta, \mathbf{I})$ is parameterized by Fusion Module parameters ϕ and depends on parameters θ of the OneFormer model and a sequence of frames $\mathbf{I}$. The parameters θ are updated by backpropagation through adaptive gradients. During the training process, the Fusion module parameters ϕ and the OneFormer parameters θ are optimized jointly. The goal is to minimize multicomponent segmentation loss $\mathcal{L}_{segm}(\theta, \mathbf{I})$ over all ground-truth sequences $\mathbf{R}_{all}$.

$$\min_{\theta,\phi} \sum_{\mathbf{I}\in\mathbf{R}_{all}} \mathcal{L}_{segm}(\theta - \alpha\nabla_\theta \mathcal{L}_{fusion}(\phi, \theta, \mathbf{I}), \mathbf{I}). \tag{1}$$

The segmentation loss function is the original OneFormer loss function [35] without the contrastive loss term. Thus,

$$\mathcal{L}_{segm} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice}, \tag{2}$$

where, $\mathcal{L}_{cls}$ – cross-entropy loss for class prediction, binary cross-entropy ($\mathcal{L}_{bce}$) and dice loss ($\mathcal{L}_{dice}$) are controlling mask predictions.

Navigation Image Sequences The SegmATRon model is trained on offline data of image sequences that were collected in a simulation environment. However, during navigation, sequences of images appear online, and the frames' content depends on the actions chosen by the agent. Therefore, the SegmATRon can use a buffer storing frame sequences that are updated interactively. Two parameters control the buffer properties: the number of images and the sequence order. In our experiments, we consider buffers of different lengths: 2 images, 3 images and 5 images. These sizes correspond to 1, 2, and 4 additional frames, i.e. agent steps, used to refine semantic segmentation. Besides, we conduct experiments with different image sequence orders in the frame buffer. We call the image sequence order “backward” when a sequence of frames $\{I_t, I_{t-1}, ..., I_{t-n}\}$ is used for prediction at step t, where n is the number of additional frames that the SegmATRon (B) uses. In the case of "forward" image sequence order, at time $t+n$ the SegmATRon (F) predicts frame I_t, using "future" frames relatively to frame I_t, i.e. $\{I_t, I_{t+1}, ..., I_{t+n}\}$. Thus, the semantic map is updated with a delay, but the SegmATRon (F) model can use frames in which the goal objects are better viewed. When a goal is observed, the agent navigates toward it, so the “forward” images could have more information about the goal than the “backward” images.

F. SEMANTIC MAP ACCUMULATION AND FILTERING

Learning-based semantic segmentation predictions sometimes have noises, especially in far objects. Projecting these noises onto the map causes semantic mapping outliers and leads to reaching a false goal and an unsuccessful finish. To prevent this, we implement semantic map filtering consisting of erosion, dilation, map accumulation, and fading. A scheme of semantic map filtering is shown in Fig. 5.

At each step t, a local semantic map L_t is created using pose and depth observation with the SegmATRon-predicted semantic mask. First, an erosion is applied to the local map to filter out semantic segmentation outliers. Next, to return the initial size to the target objects, dilation is applied as a convolution. After that, the local map is fused with the global map M_{t-1} to obtain the updated global map M_t. The fusion is implemented as accumulation and fading. The global map cells accumulate information about the presence of a goal object on the local map at every step. The global map values in the cells containing a goal object in the local map are

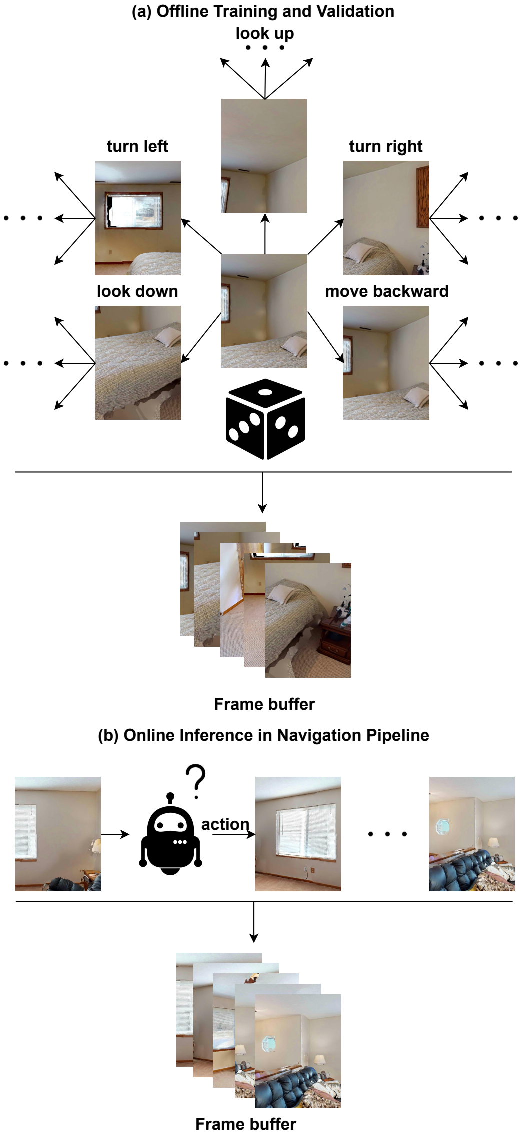


FIGURE 6. Frame sequence collection during training of the interactive segmentation model SegmATRon and inference in the navigation pipeline. The number of additional frames during both training and inference remains the same. (a) During offline training and validation, additional frames are randomly selected from a predetermined list of actions: turn right, turn left, look up, look down, and move backward. (b) During inference in the navigation pipeline, the next frame is determined using navigation skills. The frame buffer stores the history of the agent's observations while navigating through the environment.

increased by 1. If a goal object exists on the global map but is absent on the local map, then it is gradually erased from the global map using the fading mechanism. The global map values in the cells not containing a goal object on the local map are multiplied by a decay coefficient $\alpha < 1$. The global map values in the cells not covered by a local map are not changed. A cell is considered a goal object cell if its value in the global map exceeds some predefined threshold T. A formal description of the semantic map filtering is shown below:

$$L_t = dilation_k(erosion_k(L_t)), \tag{3}$$

$$\begin{aligned} M_t = (M_{t-1} + L_t) \cdot (L_t + \alpha(1 - L_t)C_t \\ +(1 - L_t)(1 - C_t)), \end{aligned} \tag{4}$$

where C_t is the coverage mask of the local map L_t.

V. EXPERIMENTS

A. EXPERIMENTAL SETUP

Navigation. We validate our navigation pipeline in the photo-realistic Habitat simulator [37]. We use validation scenes from the Habitat Matterport3D semantics (HM3DSem) v0.2 dataset [6] and select 108 episodes with the following distribution of goal categories: 18 episodes with potted plants, 15 episodes with sofas, 25 episodes with chairs, 17 episodes with TVs, 17 episodes with beds, and 16 episodes with toilets. The validation scenes were never seen during the training phase. We use the environment configuration from Habitat Challenge 2023 [38], except with slight changes. The use of adaptive gradients during inference requires more computational sources for the SegmATRon method. Therefore, we increase the maximum amount of time per episode up to 1500 seconds and fix the maximum number of steps to 500. We conduct experiments on a server with 1 Nvidia Tesla V100 GPU. We repeat each validation navigation experiment 5 times and report the mean value of standard navigation metrics: Success rate, SPL, and SoftSPL as they defined in Habitat Challenge 2023 [38].

Semantic maps representation. We assess the quality of semantic map representation for different methods using the same observation dataset. We collect this dataset by recording observations during navigation with the SkillTron approach with the SegmATRon (B) (1 Step) semantic method. At each step, we save the agent's position, the tilt of its head, the depth map, GT semantics, RGB observation, and SegmATRon (B) (1 Step) predictions of a semantic mask. Then, we make predictions on the saved image sequence using different versions of the SegmATRon approach and various sets of hyper-parameters for semantic map construction. This approach allows us to distinguish between the navigation performance and the quality of semantic map representation. We evaluate the quality of semantic map representation using metrics of Closeness-sigmoid, IoU described in Section V-B.

Interactive segmentation model training. We train SegmATRon on offline data collected in HM3DSem v0.2 dataset [6]. We collect a training dataset in 1160 random points of train scenes and a validation dataset in 144 random points of validation scenes. We make sure to include some random viewpoints of goal categories in our datasets. The image sequences for the SegmATRon models training are obtained by considering all possible combinations of 4 actions made from starting random points. We consider the following set of actions: turn left, turn right, look up, look down, and move backward. By moving backward the agent can observe a scene from a more distant point of view. The tilt angle for look up and look down action is fixed to 30°. In

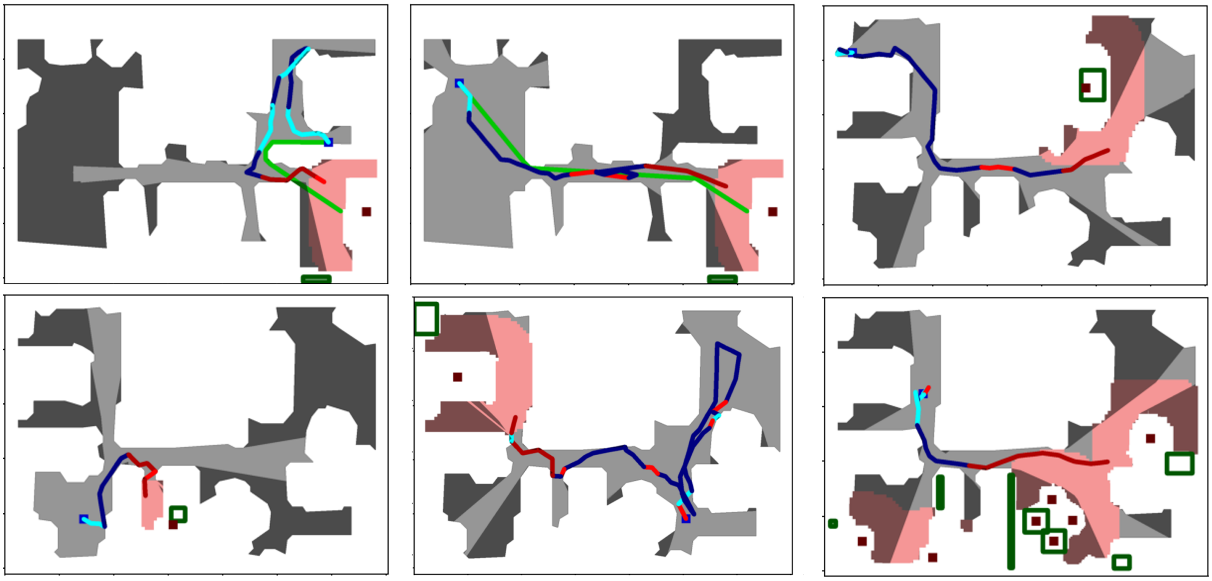


FIGURE 7. Illustration of agent trajectories in a simulated environment, utilizing the proposed SkillTron method. Dark blue and blue paths represent Classical and RL Exploration skills, respectively, while dark red and red paths depict Classical and RL GoalReacher skills. The blue square serves as the starting point, and the red region is the area where the goal target is deemed to have been reached.

our experiments, we consider turn angles of 15°and 30°. For both values of a turn angle, we collect datasets starting from the same root points. Since during the training and validation process the image sequences are chosen randomly from available combinations, we use the same sequence of weights for the SegmATRon (B) and SegmATRon (F) approaches.

In our experiments, we use a custom mapping of 1624 original categories HM3DSem v0.2 dataset [6] to 150 categories of ADE20k [39]. Thus, we can efficiently fine-tune our segmentation models starting from weights trained on the rich semantics of ADE20k [39] without pseudo-labeling. We consider this mapping as ground truth during the training and validation process of SegmATRon. Additionally, we use this mapping to compute ground-truth semantic maps to assess the quality of semantic map representation.

Figure 6 illustrates the frame buffer collection during the SegmATRon offline training process and online semantic segmentation inference in the navigation pipeline.

Baselines. As the main baseline for our approach, we consider the SkillFusion [3] method with some modifications: the classical pipeline is implemented using the PONI algorithm instead of FBE, and the semantic segmentation module is based on the OneFormer model [35] with Swin-L backbone. The SkillFusion approach with such configuration was the winner of Habitat Challenge 2023 [38]. We distinguish the role of interactive semantic map representation by considering the single-frame baseline, i.e. the OneFormer model, that was fine-tuned on the same datasets as the SegmATRon model. Also, we compare with the other methods from the Habitat challenge 2023 public leaderboard and previous state-of-the-art method for the indoor Object Goal Navigation task (refer to the Section V-C).

B. SEMANTIC MAPS

We estimate the quality of semantic maps using two metrics: Closeness-Sigmoid and Intersection over Union (IoU). The Closeness-Sigmoid metric is calculated as an average distance from the predicted semantic map cells to the closest cell of the ground truth semantic map, followed by a sigmoid function. The resulting metric has a value range between 0 and 1. If there are no predicted map cells, the metric is considered 1. Also, we estimate False Positive Rate (FPR) and False Negative Rate (FNR) metrics. The FPR value is part of episodes where the predicted map contains a spurious goal object, and the ground truth map contains no goal objects. The FNR value is part of episodes where the ground-truth semantic map contains goal objects and the predicted semantic map is empty.

First, we choose the optimal filtering hyperparameters: decay coefficient α and threshold T. For this purpose, we test four pairs (α, T) with SegmATRon (B) (1 Step). The results of the tests are shown in Table 1. According to all the metric values, we choose $\alpha = 0.9$ and $T = 2$ for our SegmATRon experiments. Next, we test different SegmATRon approaches with the chosen filtering hyperparameters. The results of the tests are shown in Table 2. According to the both Closeness-Sigmoid and IoU metrics, the best approach is the SegmATRon (B) with four steps - this approach reaches Closeness-Sigmoid 0.175 and IoU 0.446. The results with two-step SegmATRon (F) are relatively close: Closeness-Sigmoid 0.184 and IoU 0.415.

We also compare the semantic map quality using SegmATRon as a semantic method and the single-frame baseline approach, OneFormer, trained on the same dataset as SegmATRon. The OneFormer method is one of the current

TABLE 1. Semantic map quality with respect to filtering parameters

Decay coef. α	Threshold T	Closeness-sigmoid	IoU	FPR	FNR
0.8	1	0.366	0.365	0.23	0.02
0.9	1	0.392	0.358	0.24	**0.01**
0.8	2	**0.240**	0.371	**0.12**	0.03
0.9	2	0.272	**0.395**	0.13	0.02

TABLE 2. Semantic map quality with different SegmATRon approaches

Semantic method	Number of steps	Closeness-sigmoid	IoU	FPR	FNR
OneFormer	Single Frame	0.226	**0.449**	**0.06**	0.08
SegmATRon (B)	1	0.272	0.395	0.13	0.02
SegmATRon (B)	2	0.192	0.434	0.10	**0.01**
SegmATRon (B)	4	**0.175**	0.446	0.09	0.03
SegmATRon (F)	1	0.217	**0.418**	0.11	**0.02**
SegmATRon (F)	2	**0.184**	0.415	**0.08**	**0.02**
SegmATRon (F)	4	0.213	0.371	0.09	0.05

TABLE 3. Performance of SkillTron as compared to the baselines on the HM3DSem v0.2 dataset.

Method	Exploration	Goalreacher	Semantic method	Success	SPL	SoftSPL
DD-PPO [19]	-	-	-	0.07	0.04	0.28
Aux-RL [21]	-	-	-	0.18	0.10	0.31
Host_74441_Team[3]	-	-	-	0.12	0.05	0.27
ICanFly[3]	-	-	-	0.43	0.26	**0.37**
SkillFusion	FBE+RL	Classic+RL	OneFormer	0.55	0.26	0.34
SkillTron	**PONI + RL**	**Classic + RL**	**SegmATRon(B)(2Steps)**	**0.59**	**0.28**	0.36

TABLE 4. Ablation study. Number of steps and Turn Angle Values.

Method	Semantic method	Number of steps	Turn Angle	Success	SPL	SoftSPL
SkillTron	OneFormer	-	30°	0.57	0.28	0.35
SkillTron	*SegmATRon(B)*	1	30°	0.58	**0.29**	0.35
SkillTron	**SegmATRon(B)**	**2**	**30°**	**0.59**	0.28	**0.36**
SkillTron	SegmATRon (B)	4	30°	0.57	0.27	0.35
SkillTron	SegmATRon (B)	1	15°	0.51	0.25	0.33
SkillTron	SegmATRon (B)	2	15°	0.49	0.24	0.31
SkillTron	SegmATRon (B)	4	15°	0.50	0.25	0.32

state-of-the-art approaches for semantic segmentation. Table 2 shows that the use of the SegmATRon (B) with 2 and 4 steps significantly improves the quality of semantic map construction according to the Closeness-sigmoid metric. The values of the IoU metric turn out to be close for both approaches. The semantic map built using the SegmATRon contains a significantly smaller number of false negative goal predictions. Thus, the total number of episodes in which the goal was mispredicted is lower when using the SegmATRon as the semantic segmentation method.

C. NAVIGATION WITH DIFFERENT SEMANTIC MAPS

Comparison with different baselines. The SkillTron approach significantly outperforms various state-of-the-art methods listed on the public leaderboard of Habitat Challenge 2023 Test Standard Phase [40], [38] (see Table 3). There is no public code available for the ICanFly and Host_74441_Team approaches, therefore we report the navigation metrics based on its performance on the Habitat Challenge 2023 Test Standard dataset [38]. As one can see from Table 3, the SkillTron surpasses the state-of-the-art approach SkillFusion [3] by a considerable margin of 4% of success rate, 2% of SPL and 2% of SoftSPL metrics. The SkillTron significantly outperforms previous state-of-the-art methods for the indoor Object Goal Navigation task such as DDPO [19]and Aux-RL [21].

The interactive semantic map representation plays an important role in the performance of the SkillTron method. Our best interactive segmentation network uses two additional frames and a backward image sequence. The SkillTron method with interactive semantic map representation shows the increase of 2% of the Success and 1% of SoftSPL metrics compared to the baseline SkillTron using OneFormer as the semantic segmentation network.

Ablation on the number of steps. Long image sequences carry more information about the environment. We vary the

[3] We report metrics from the public leaderboard of Habitat Challenge 2023 Test Standard Phase [40], [38] due to the absence of public code implementation.

number of additional frames used to predict a segmentation mask. We consider the backward image sequence order for these experiments and the turn angle of 30°. Table 4 shows the navigation metrics for SegmATRon (B) with a different number of steps, i.e. additional frames. The SegmATRon (B) (2 Steps) demonstrates the increase of success rate and SoftSPL metrics compared to the SegmATRon (B) (1 Steps). However, the SegmATRon (B) (4 Steps) shows a decrease in navigation metrics compared to the SegmATRon (B) (2 Steps). These results are in agreement with the dependence of semantic maps quality metrics on the number of steps (see Section V-B) for 1 and 2 additional frames. The 4 additional frames may not be optimal for navigation, since the processing of longer image sequences slows the navigation pipeline.

Ablation on the turn angle values. The continuity of view during navigation is determined by the agent's turn angle. We conduct an ablation study to investigate if a smaller turn angle would increase the performance of the Fusion module of the interactive segmentation network. We consider 15° a half of turn angle in the configuration of discrete action space of Habitat Challenge 2023 [38]. We retrain learning-based skills in the new action space and the interactive segmentation network. The 15° turn angle is less effective for navigation than 30° (see Table 4) for all considered number of steps. It can be partially since the smaller turn angle requires more steps during the exploration phase of object goal navigation.

VI. CONCLUSION

In this paper, we propose a new visual navigation approach called SkillTron, which utilizes a two-level interactive semantic map representation, as well as fusing the Exploration and GoalReacher skills of the robot. To construct a one-shot map level, we examine in detail the neural network method, which adjusts the weights of the segmentation model based on the predicted values of fusion loss during inference on a regular (backward) or delayed (forward) image sequence. We demonstrate that the backward interaction mode provides a more accurate construction of a 2D accumulated semantic map, which is then used for navigation. It is shown that the proposed combination of an RL-based navigation pipeline and a classic modular approach using learnable modules outperforms existing state-of-the-art approaches in indoor environments from the Habitat simulator.

As limitations, it should be noted that the chosen semantic segmentation network is resource-demanding, and the proposed visual navigation approach has only been tested in simulation environments. Further directions for the development of the proposed approaches could include studying the transfer of the visual navigation method to a real robot, using other more compact basic models of semantic segmentation and image sequence fusion to form a one-shot representation of a semantic map. Considering interactive segmentation as a separate robot skill with a learned action policy is also of interest for future work.

REFERENCES

[1] K. Muravyev, A. Bokovoy, and K. Yakovlev, "Enhancing exploration algorithms for navigation with visual slam," in *Proc. RCAI*, Taganrog, Russia, 2021, pp. 197–212.

[2] D. Gordon, A. Kadian, D. Parikh, J. Hoffman, and D. Batra, "Splitnet: Sim2sim and task2task transfer for embodied visual navigation," in *Proc. IEEE/CVF ICCV*, Seoul, Republic of Korea, 2019, pp. 1022–1031.

[3] A. Staroverov, K. Muravyev, K. Yakovlev, and A. I. Panov, "Skill fusion in hybrid robotic framework for visual object goal navigation," *Robotics*, vol. 12, no. 4, pp. 104–118, July 2023, doi: 10.3390/robotics12040104.

[4] P. Ewen, A. Li, Y. Chen, S. Hong, and R. Vasudevan, "These maps are made for walking: Real-time terrain property estimation for mobile robots," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 7083–7090, July 2022, doi: 10.1109/LRA.2022.3180439.

[5] S. Yang, Y. Huang, and S. Scherer, "Semantic 3d occupancy mapping through efficient high order crfs," in *Proc. IEEE/RSJ IROS*, Vancouver, BC, Canada, 2017, pp. 590–597.

[6] K. Yadav, R. Ramrakhya, S. K. Ramakrishnan, T. Gervet, J. Turner, A. Gokaslan, N. Maestre, A. X. Chang, D. Batra, M. Savva *et al.*, "Habitat-matterport 3d semantics dataset," in *Proc. IEEE/CVF CVPR*, Vancouver, BC, Canada, 2023, pp. 4927–4936.

[7] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv*, Dec. 2017, doi: 10.48550/arXiv.1712.05474.

[8] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, M. Anvari, M. Hwang, M. Sharma, A. Aydin, D. Bansal, S. Hunter, K.-Y. Kim, A. Lou, C. R. Matthews, I. Villa-Renteria, J. H. Tang, C. Tang, F. Xia, S. Savarese, H. Gweon, K. Liu, J. Wu, and L. Fei-Fei, "Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation," in *Proc. CoRL*, Auckland, New Zealand, 2022, pp. 80–93.

[9] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. CoRL*, Mountain View, California, USA, 2017, pp. 1–16.

[10] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*, Zurich, Switzerland, 2017, pp. 621–635.

[11] NVIDIA, "Isaac sim," 2021, accessed on: February, 18, 2023. [Online]. Available: https://developer.nvidia.com/isaac-sim

[12] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation," in *Proc. IEEE/CVF CVPR*, Vancouver, BC, Canada, 2023, pp. 23 171–23 181.

[13] D. S. Chaplot, D. Gandhi, A. Gupta, and R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *Proc. NeurIPS*, Online Conference, Canada, 2020, pp. 4247–4258.

[14] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "Poni: Potential functions for objectgoal navigation with interaction-free learning," in *Proc. IEEE/CVF CVPR*, New Orleans, Louisiana, USA, 2022, pp. 18 890–18 900.

[15] A. Asgharivaskasi and N. Atanasov, "Semantic octree mapping and shannon mutual information computation for robot exploration," *IEEE T-RO*, vol. 39, no. 3, pp. 1910–1928, June 2023, doi: 10.1109/TRO.2023.3245986.

[16] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: an open-source library for real-time metric-semantic localization and mapping," in *Proc. IEEE ICRA*, Paris, France, 2020, pp. 1689–1696.

[17] R. Zurbrügg, H. Blum, C. Cadena, R. Siegwart, and L. Schmid, "Embodied active domain adaptation for semantic segmentation via informative path planning," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 8691–8698, Oct. 2022, doi: 10.1109/LRA.2022.3188901.

[18] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proc, IEEE CIRA*, Monterey, Calif., USA, 1997, pp. 146–151.

[19] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," in *Proc. ICLR*, Addis Ababa, Ethiopia, 2020, accessed on: February, 18, 2023. [Online]. Available: https://openreview.net/pdf?id=H1gX8C4YPr

[20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint*, July 2017, doi: 10.48550/arXiv.1707.06347.

[21] J. Ye, D. Batra, A. Das, and E. Wijmans, "Auxiliary tasks and exploration enable objectgoal navigation," in *Proc. IEEE/CVF ICCV*, Montreal, BC, Canada, 2021, pp. 16 117–16 126.

[22] W. Ding, N. Majcherczyk, M. Deshpande, X. Qi, D. Zhao, R. Madhivanan, and A. Sen, "Learning to view: Decision transformers for active object detection," *arXiv preprinnt*, Jan. 2023, doi: 10.48550/arXiv.2301.09544.

[23] J. Yang, Z. Ren, M. Xu, X. Chen, D. J. Crandall, D. Parikh, and D. Batra, "Embodied amodal recognition: Learning to move to perceive objects," in *Proc. IEEE/CVF ICCV*, Seoul, South Korea, 2019, pp. 2040–2050.

[24] G. Chaudhary, L. Behera, and T. Sandhan, "Active perception system for enhanced visual signal recovery using deep reinforcement learning," in *Proc. IEEE ICASSP*, Rhodes Island, Greece, 2023, pp. 1–5.

[25] P. Hoseini, S. K. Paul, M. Nicolescu, and M. Nicolescu, "A one-shot next best view system for active object recognition," *Applied Intelligence*, vol. 52, no. 5, pp. 5290–5309, Aug. 2022, doi: 10.1007/s10489-021-02657-z.

[26] J. Li and M. Bansal, "Improving vision-and-language navigation by generating future-view image semantics," in *Proc. IEEE/CVF CVPR*, Vancouver, BC, Canada, 2023, pp. 10 803–10 812.

[27] K. P. Singh, L. Weihs, A. Herrasti, A. Kembhavi, and R. Mottaghi, "Ask4help: Learning to leverage an expert for embodied tasks," in *Proc. NeurIPS*, New Orleans, Louisiana, USA, 2022, accessed on: February, 18, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/674ad201bc8fa74b3c9979230aa0c63b-Paper-Conference.pdf

[28] Z. Fang, A. Jain, G. Sarch, A. W. Harley, and K. Fragkiadaki, "Move to see better: Towards self-supervised amodal object detection," *arXiv preprint*, 2020, doi: 10.48550/arXiv.2012.00057.

[29] Y. Jing and T. Kong, "Learning to explore informative trajectories and samples for embodied perception," *arXiv preprint*, 2023, doi: 10.48550/arXiv.2303.10936.

[30] K. Kotar and R. Mottaghi, "Interactron: Embodied adaptive object detection," in *Proc. IEEE/CVF CVPR*, New Orleans, Louisiana, USA, 2022, pp. 14 860–14 869.

[31] T. Zemskova, M. Kichik, D. Yudin, A. Staroverov, and A. Panov, "Segmatron: Embodied adaptive semantic segmentation for indoor environment," *arXiv preprint*, 2023, doi: 10.48550/arXiv.2310.12031.

[32] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects," *arXiv preprint*, 2020, doi: 10.48550/arXiv.2006.13171.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Munich, Germany, 2015, pp. 234–241.

[34] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 4, no. 2, pp. 100–107, July 1968, doi: 10.1109/TSSC.1968.300136.

[35] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "Oneformer: One transformer to rule universal image segmentation," in *Proc. IEEE/CVF CVPR*, Vancouver, BC, Canada, 2023, pp. 2989–2998.

[36] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, Glasgow, United Kingdom, 2020, pp. 213–229.

[37] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A Platform for Embodied AI Research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

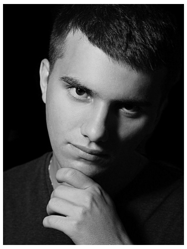
[38] K. Yadav, J. Krantz, R. Ramrakhya, S. K. Ramakrishnan, J. Yang, A. Wang, J. Turner, A. Gokaslan, V.-P. Berges, R. Mootaghi, O. Maksymets, A. X. Chang, M. Savva, A. Clegg, D. S. Chaplot, and D. Batra, "Habitat challenge 2023," 2023, accessed on: February, 18, 2023. [Online]. Available: https://aihabitat.org/challenge/2023/

[39] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 302–321, Mar. 2019, doi: 10.1007/s11263-018-1140-0.

[40] K. Yadav, J. Krantz, R. Ramrakhya, S. K. Ramakrishnan, J. Yang, A. Wang, J. Turner, A. Gokaslan, V.-P. Berges, R. Mootaghi, O. Maksymets, A. X. Chang, M. Savva, A. Clegg, D. S. Chaplot, and D. Batra, "Habitat challenge 2023 public leaderboard," 2023, accessed on: February, 18, 2023. [Online]. Available: https://eval.ai/web/challenges/challenge-page/1992/leaderboard/4705

TATIANA ZEMSKOVA received an M.S. degree in Applied Mathematics and Computer Science from the Moscow Institute of Physics and Technology, Moscow, Russia, 2023 and an M.S. degree in Engineering from Ecole Polytechnique, Palaiseau, France, 2023. She is currently pursuing a Ph.D. degree in computer science at the Moscow Institute of Physics and Technology, Moscow, Russia.

From 2023 to the present, she has been working as an Engineer at the Intelligent Transport Laboratory at the Moscow Institute of Physics and Technology, Moscow, Russia. Her research interests include computer vision, embodied AI and robotic systems.



ALEKSEI STAROVEROV received an M.S. degree from Bauman Moscow State Technical University, Moscow, Russia in 2019. He is currently pursuing a Ph.D. degree in computer science at the Moscow Institute of Physics and Technology, Moscow, Russia. His research thesis involves the methods and algorithms for the automatic determination of subgoals in a reinforcement learning problem for robotic systems.

From 2022 to the present, he has been working as a Researcher at the Artificial Intelligence Research Institute, Moscow, Russia. His research interests include reinforcement learning, deep learning, and robotic systems.

KIRILL MURAVYEV received the M.S. degree in computer science from Moscow Institute of Physics and Technology, Moscow, Russia, in 2021. He is currently pursuing the Ph.D. degree in computer science in Federal Research Center for Computer Sciences and Control of Russian Academy of Sciences, under the supervision of K. Yakovlev.

He is currently a Junior Researcher with the Federal Research Center for Computer Sciences and Control of Russian Academy of Sciences. His research interests include SLAM, map construction methods, and mobile robots navigation.

DMITRY A. YUDIN received the engineering diploma in automation of technological processes and production in 2010 and the Ph.D. degree in computer science from the Belgorod State Technological University (BSTU) named after V.G. Shukhov, Belgorod, Russia in 2014.

From 2009 to 2019, he was a Researcher and Assistant Professor with Technical Cybernetics Department at BSTU n.a. V.G. Shukhov. Since 2019, he has been the head of the Intelligent Transport Laboratory at the Moscow Institute of Physics and Technology, Moscow, Russia. Since 2021, he has been a Senior Researcher at AIRI (Artificial Intelligence Research Institute), Moscow, Russia. He is the author more than 100 articles. His research interests include computer vision, deep learning, and robotics.

ALEKSANDR I. PANOV earned an M.S. in Computer Science from the Moscow Institute of Physics and Technology, Moscow, Russia, 2011 and a Ph.D. in Theoretical Computer Science from the Institute for Systems Analysis, Moscow, Russia, in 2015.

Since 2010, he has been a research fellow with the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia. Since 2018, he has headed the Cognitive Dynamic System Laboratory at the Moscow Institute of Physics and Technology, Moscow, Russia. He authored three books and more than 100 research papers. In 2021, he joined the research group on Neurosymbolic Integration at the Artificial Intelligence Research Institute, Moscow, Russia. His academic focus areas include behavior planning, reinforcement learning, embodied AI, and cognitive robotics.