Habitat-Matterport 3D Semantics Dataset

Karmesh Yadav[1*], Ram Ramrakhya[2*], Santhosh Kumar Ramakrishnan[3*],
Theo Gervet[6], John Turner[1], Aaron Gokaslan[4], Noah Maestre[1], Angel Xuan Chang[5],
Dhruv Batra[1,2], Manolis Savva[5], Alexander William Clegg[1†], Devendra Singh Chaplot[1†]
[1]Meta AI [2]Georgia Tech [3]UT Austin
[4]Cornell University [5]Simon Fraser University [6]Carnegie Mellon University

Abstract

*We present the Habitat-Matterport 3D Semantics (*HM3DSEM*) dataset.* HM3DSEM *is the largest dataset of 3D real-world spaces with densely annotated semantics that is currently available to the academic community. It consists of 142,646 object instance annotations across 216 3D spaces and 3,100 rooms within those spaces. The scale, quality, and diversity of object annotations far exceed those of prior datasets. A key difference setting apart* HM3DSEM *from other datasets is the use of texture information to annotate pixel-accurate object boundaries. We demonstrate the effectiveness of* HM3DSEM *dataset for the Object Goal Navigation task using different methods. Policies trained using* HM3DSEM *perform outperform those trained on prior datasets. Introduction of* HM3DSEM *in the Habitat ObjectNav Challenge lead to an increase in participation from 400 submissions in 2021 to 1022 submissions in 2022. Project page:* https://aihabitat.org/datasets/hm3d-semantics/

1. Introduction

Over the recent past, work on acquiring and semantically annotating datasets of real-world spaces has significantly accelerated research into embodied AI agents that can perceive, navigate and interact with realistic indoor scenes [1–5]. However, the acquisition of such datasets at scale is a laborious process. HM3D [5] which is one of the largest available datasets with 1000 high-quality and complete indoor space reconstructions, reportedly required 800+ hours of human effort to carry out mainly data curation and verification of 3D reconstructions. Moreover, dense semantic annotation of such acquired spaces remains incredibly challenging.

We present the Habitat-Matterport 3D Dataset Semantics (HM3DSEM). This dataset provides a dense semantic annotation 'layer' augmenting the spaces from the original HM3D dataset. This semantic 'layer' is implemented as a set of textures that encode object instance semantics and cluster objects into distinct rooms. The semantics include architectural elements (walls, floors, ceilings), large objects (furniture, appliances etc.), as well as 'stuff' categories (aggregations of smaller items such as books on bookcases). This semantic instance information is specified in the semantic texture layer, providing pixel-accurate correspondences to the original acquired RGB surface texture and underlying geometry of the objects.

The HM3DSEM dataset currently contains annotations for 142,646 object instances distributed across 216 spaces and 3,100 rooms within those spaces. Figure 1 shows some examples of the semantic annotations from the HM3DSEM dataset. The achieved scale is larger than prior work (2.8x relative to Matterport3D [6] (MP3D) and 2.1x relative to ARKitScenes [7] in terms of total number of object instances). We demonstrate the usefulness of HM3DSEM on the ObjectGoal navigation task. Training on HM3DSEM results in higher cross-dataset generalization performance. Surprisingly, the policies trained on HM3DSEM perform better on average across scene datasets compared to training on the datasets themselves. We also show that increasing the size of training datasets improve the navigation performance. These results highlight the importance of improving the quality and scale of 3D datasets with dense semantic annotations for improving downstream embodied AI task performance.

2. Related Work

3D reconstruction datasets with semantics. There is a relatively small number of prior works that focus on semantically annotated 3D interior spaces acquired from the real world. Collecting, reconstructing, and annotating such data at scale is a significant effort that requires complex pipelines and annotation tools. Earlier work has therefore focused on scenes at the scale of single rooms. For example, ScanNet [8] provided 707 typically room-scale reconstructions annotated with object semantic instances through labeling

[*] Equal Contribution, Correspondence: karmeshyadav@meta.com
[†] Equal Contribution

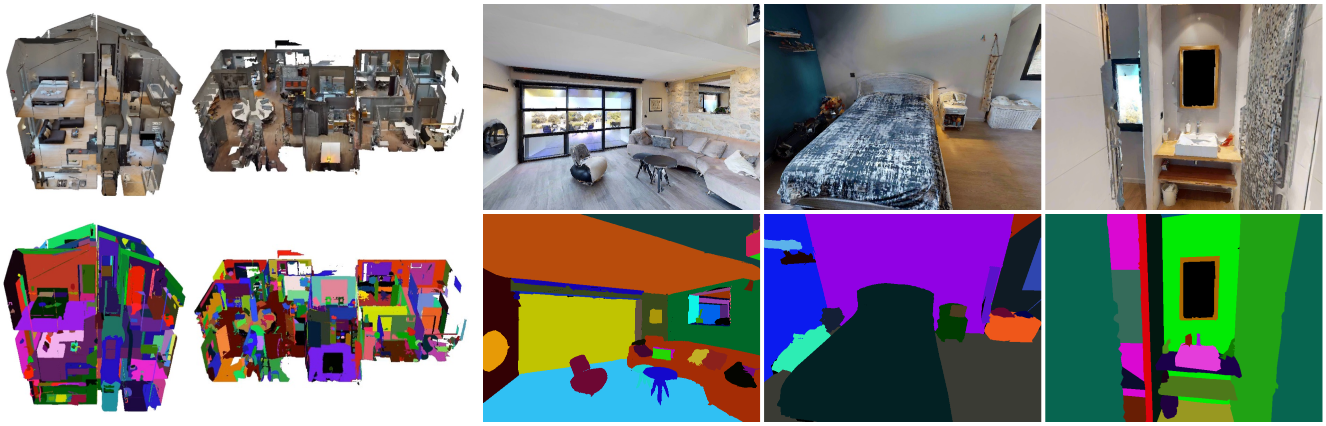

Figure 1. Habitat-Matterport 3D Semantics (HM3DSEM) provides the largest dataset of real-world spaces with densely annotated semantics. High-fidelity textured 3D mesh reconstructions are labeled with precise instance-level object semantics, indicated by distinct colors.

of 3D mesh segments constructed using an unsupervised segmentation algorithm. Followup work by Wald et al. [9] adopted a similar approach and also targeted room-sized scenes. Most recently, ARKitScenes [7] contributed scans of 1661 room-scale scenes but only provides bounding box annotations for object instances.

Prominent prior works on building-scale datasets with semantic annotation are Matterport3D [6], a subset of Gibson by Armeni et al. [10], and the Replica [11] dataset. The first uses the same methodology as ScanNet (labeling of 3D mesh segments), while the second provides human-verified object instance annotations created by back-projecting 2D semantic segmentation masks. The third provides high-quality mesh vertex-level object instance labels but only contains 18 scenes. Building on top of HM3D, which consists of over 1,000 diverse environments from around the world, HM3DSEM provides detailed texture-level semantic annotations for building-scale reconstructions.

Synthetic 3D scene datasets. The use of synthetic 3D datasets for embodied AI simulation is quite common, especially when interactive environments are desired [4, 12–14]. Due to the difficulty of modeling high-fidelity synthetic environments at scale, most existing datasets are limited in size and typically represent room-scale scenes. Some of the prior work in this space has adopted a 'teleportation' mechanism that allows an agent to immediately move from room to room through closed doors [13]. A few datasets contributed by prior work focus on larger-scale scenes that coherently represent entire residences with multiple rooms [4, 15, 16]. These datasets have a number of limitations. First, due to the difficulty in modeling a broad diversity of objects and scene layouts containing them, there is fairly limited variation in both object appearance and the spatial arrangements of the objects in the scenes. Moreover, the objects exhibit modeling biases that create a simulation-to-reality gap, and the re-use of the same object models across scenes produces the unrealistic effect of "perfect copies" of particular objects. These limitations have inspired work that attempts to tackle sim-to-real discrepancy by creating synthetic datasets that conform to scenes from the real world in terms of object appearance and spatial arrangement [4, 17–19]. However, this approach is hard to scale, and modeling biases due to the use of synthetic 3D data content creation software still remain. In contrast, we focus on scaling high-quality semantic annotations of *real* scenes acquired from a diverse set of spaces in the real world.

3. Dataset Details

The Habitat-Matterport 3D Semantics Dataset is the largest-ever human-annotated dataset of semantically-annotated 3D indoor spaces. It contains dense semantic annotations for 216 high-resolution, 3D, scanned scenes from the Habitat-Matterport 3D Dataset (HM3D). The HM3D scenes are annotated with 142,646 raw object names additionally mapped to the 40 Matterport 3D categories [6]. On average, each scene consists of 661 objects from 106 categories. This dataset is the result of over 14,200 hours of human effort for annotation and verification by 20+ annotators. The following subsections provide further details on asset formats, the annotation pipeline, and scene content statistics.

3.1. Data Format and Contents

The semantic annotations are available as a set of texture images applied to the original scene geometry from HM3D and packed into binary glTF (.glb) format. Unique hex colors differentiate each object instance and map it to a raw text string classifying the instance. These mappings are included in a metadata text file accompanying the .glb asset, which additionally labels each instance with a region ID to define object grouping by room.

[1] Human-verified subset of Gibson [20] with semantic annotations.

Dataset	Scenes	Rooms	Object instances	Objects/room	Annotation type
Replica [11]	18	≈ 25	2,843	≈ 114	vertex
Gibson (tiny[1]) [10]	35	727	2,397	≈ 3	vertex
ScanNet [8]	707	≈ 707	36,213	≈ 24	segment
3RScan [9]	478	≈ 478	43,006	≈ 29	segment
MP3D [6]	90	2,056	50,851	≈ 25	segment
ARKitScenes [7]	1,661	5,048	67,791	≈ 13	bounding box
HM3DSem (ours)	216	3,100	142,646	≈ 60	texture

Table 1. Comparison of HM3DSem to other semantically annotated indoor scene datasets. Statistics are on the publicly released portions of the corresponding datasets (does not include ScanNet or ARKitScenes hidden test sets).

Often, semantic annotations are defined per-vertex and directly embedded in the mesh geometry (e.g., ScanNet [8], Gibson [3], and MP3D [6]). However, it is not uncommon for mesh geometry discretization to insufficiently capture boundaries between objects, especially on flat surfaces such as walls, floors, and table-tops. This results in jagged inaccurate semantic boundaries, missing annotations, or requires generating an entirely new mesh with higher resolution than the original, which has implications on both rendering performance and visual alignment. For example, Figure 4 highlights the common misalignment errors between annotated and original assets from the MP3D dataset resulting from automated mesh geometry generation. In contrast, HM3DSem archival format encodes annotations directly in a set of textures compatible with the original geometry. As it is not uncommon for 3D assets, especially those derived from scanning pipelines to represent object boundaries in texture rather than geometry, this choice seemed natural. Figure 2 shows several example scenes and contrasts them against semantic annotations from Matterport3D [6], which is the most related prior dataset. The density and quality of semantic instance annotations in HM3DSem exceeds that of prior work as shown in Table 1. For additional compatibility with existing simulators, the semantic texture annotations are also baked into per-vertex colors included with the assets.

Artists were instructed to annotate architectural features such as: walls, floors, ceilings, windows, stairs, and doors as well as notable embellishments such as door and window frames, banisters, area rugs, and moulding. Instance annotations for architectural features are broken into regions at transition points such as room boundaries, doorways, and hallways to more readily classify components into regions (e.g. to semantically separate floors and ceilings as a room transitions to a hallway) as shown in Figure 4 (right). Additionally, decorative features such as pictures, posters, switches, vents, lighting fixtures, and wall art are segmented and labeled.

Furniture, appliances, and clutter objects were annotated and segmented from their surroundings whenever possible. For example, pillows and blankets are segmented individually from beds, couches, and chairs while remote controls, electronics, lamps, and art pieces are segmented from desks, tables, and consoles. In many cases, as scan resolution permits, individual clothing items, linens, and books are segmented from one another in closets and bookshelves.

3.2. Verification Process

Annotation on the scale of HM3D Semantics is not a one-way street. Roughly 640 annotator hours were allocated to iteration and error correction (about 4.5% of all annotator hours). Additional verification was done by the authors, including both qualitative manual assessment and automated programmatic checks. Even so, some errors may yet remain. Fortunately, the archival format of texture + text allows for efficient iterative improvement of the annotations.

Automated verification is essential for large scale annotation efforts. Our automated verification pipeline included, among others, the following checks:

- Text file annotations contain only colors from textures.
- Each annotation color used only once per scene.
- Text file contents conform to expected format: index, color, category name, region id.

Qualitative verification proves challenging to automate, and as such, manual validation by humans remains an important part of the annotation QA pipeline. Following delivery of the annotated assets, a manual review and iteration phase was conducted, including the following:

- Validation pass over raw text names included identification and correction of typos, consolidation of synonyms, and mapping of raw text names to the 40 canonical object classes from the MP3D dataset [6].
- Visual inspection through virtual walk-through in Habitat [4]. Verifiers checked for missing annotations, messy boundaries, annotation artifacts, over-aggregation (i.e., multiple unique instances sharing an annotation color), semantic mislabeling (e.g. "dishwasher" annotated as "washing machine"), and other common flaws.

3.3. Dataset Statistics

The 216 scenes chosen as candidates for HM3DSem annotation were selected at random from the 950 furnished HM3D scan assets. These are distributed into subsets of [145, 36, 35] scenes between [train, val, test] splits. The

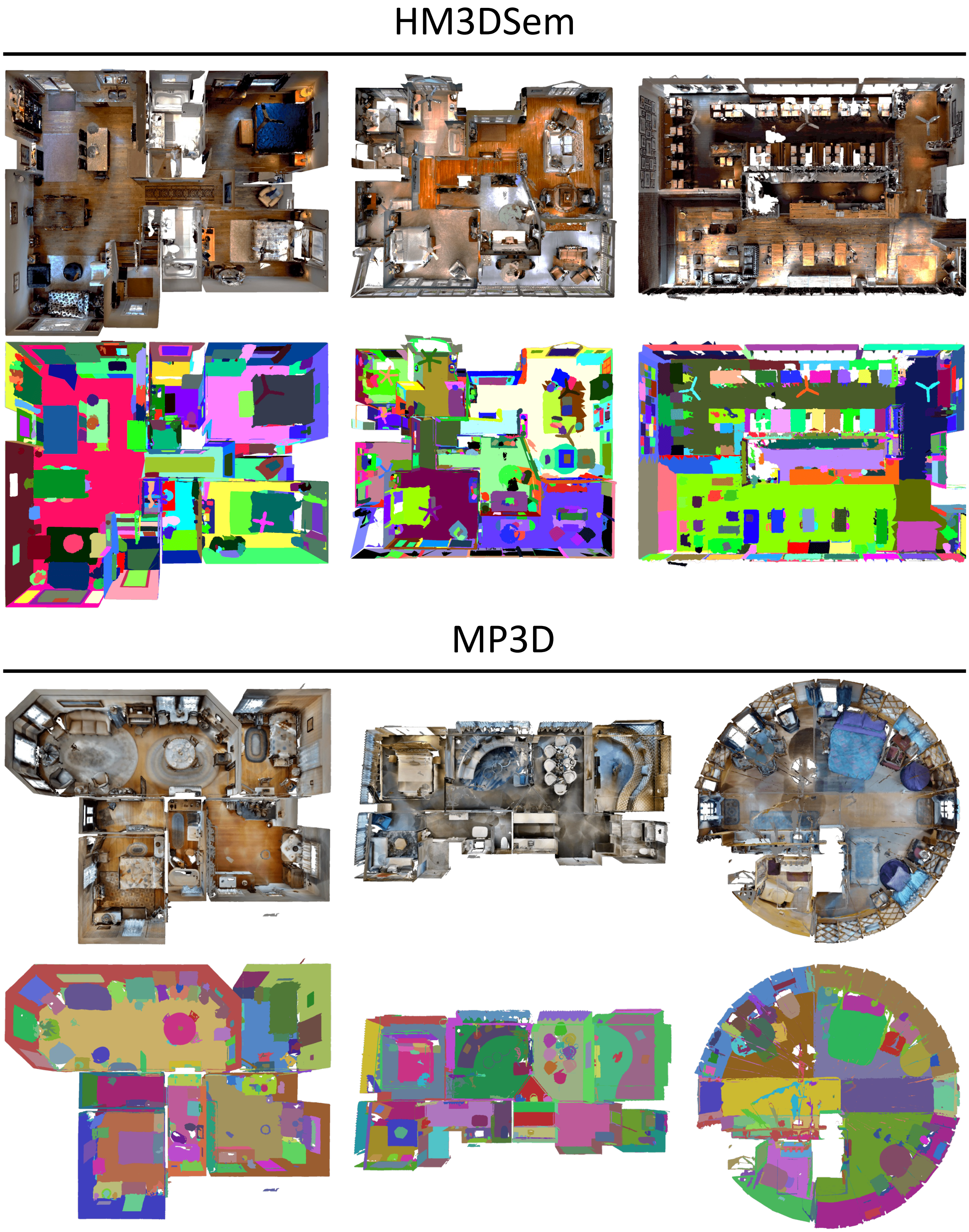

Figure 2. Qualitative examples comparing semantic annotations of scenes from HM3DSem (top) and Matterport3D [6] (bottom). The first row in each pair of rows shows a top-down view of the scene. The second row shows semantic object instances in distinct colors. The HM3DSem annotations provide a greater number of distinct object instances, as indicated also by the summary statistics in Table 1). Many paintings and other wall objects are annotated in HM3DSem (see leftmost wall in top right scene). Smaller object types such as decorative pieces on bookcases (see top left scene, leftmost corner) are also annotated. In contrast, semantic annotations from Matterport3D tend to cluster smaller objects into larger furniture pieces (e.g., items on piano at top left, and items on nightstand and bed in rightmost scene).

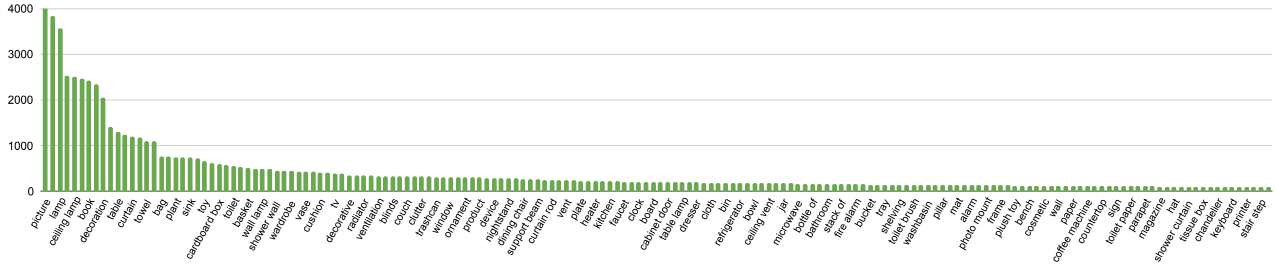

Figure 3. Histogram of most common semantic labels with 100+ instances across all scenes. Common architectural categories (e.g. floor, ceiling, wall) are not displayed.

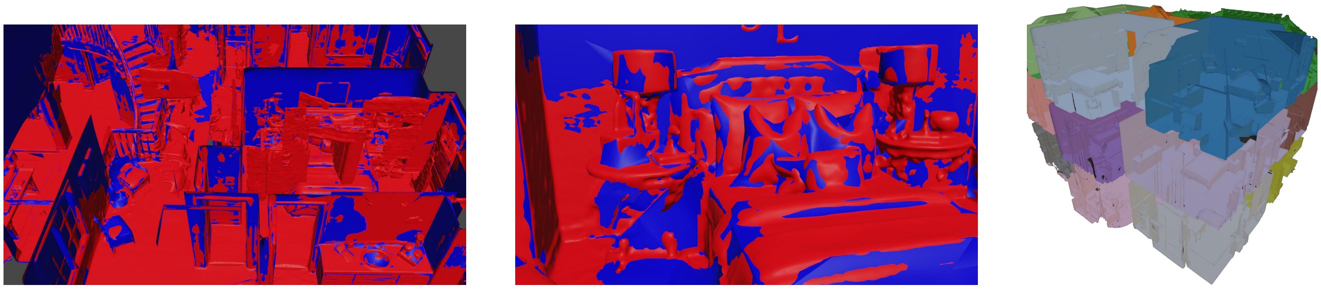

Figure 4. Examples of major (**left**) and minor (**center**) misalignment between the scanned RGB mesh (blue) and auto-generated semantic mesh (red) from the MP3D dataset [6]. **Right**: Visualization of region annotations in HM3DSEM. Each color is an aggregation of all instances mapped to a particular region.

36th val model is an example scene freely available without registration for quick inspection and automated testing of downstream dataset use cases.

An analysis of the annotation text files reveals much about the contents of the scanned environments. There are 1,625 category tags labeling the 142,646 object instances across the entire dataset, split amongst 3,100 regions. Each scene contains, on average, 106 unique categories, 660 object instances, and 14 annotated regions. The histograms in Figure 6 show the overall distribution of regions, object instances, and unique categories across all scenes in the dataset. It is worth noting that because annotators were given freedom when defining category tags, many synonymous tags are present in the final dataset.

Of the 142,646 object instances present in the dataset, 34,368 are either labeled as “unknown” or belong to architectural categories, such as “wall”, “door”, “ceiling”, etc, leaving approximately 108,278 annotated object instances. Those categories with 100 or more instances throughout the dataset are shown in Figure 3.

Each annotated region contains on average, 46 object instances and 20 unique categories. Further statistical observations were made using a region labeling heuristics where proposed region/room labels were derived from object category-inferred proposals. For example the presence of a “bed” instance implies that the containing region is a “bedroom” and a “toilet” implies a “bathroom”. See the supplemental material for more details and source data sheets.

Given regions labeled using these heuristics, we can investigate the prevalence of individual room types within the scenes and cluster them by their expected contents. Figure 5 shows a histogram of common room types counted per-scene. From these statistics we can see that:

- More than 30% of scenes are larger residences with 4+ bedrooms and bathrooms.
- Many scenes have more than 1 kitchen, possibly indicating multi-family homes or multiple individual living units packed into a single scene.
- A small set of very large scenes have 5+ regions labeled as offices and living rooms.

Further analysis of the raw data reveals:

- 14 scenes lacked any heuristically labeled "bedroom" regions. These scenes were all visually verified to be commercial spaces such as offices, restaurants, or stores.
- 7 scenes lacked any “bathroom” regions. These were also visually verified to be non-residential spaces (a subset of the 14 which lacked "bedroom" labels).
- 25 scenes contained “garage” regions. These were visually verified to all contain garages.
- 13 scenes lacked any “kitchen” regions. 5 of these were commercial spaces that also lacked “bedroom” labels, while 5 of the remaining 6 were hotel rooms or suites. The final scene contained a kitchen through visual inspection, however the modern design lacked most obvious appliances and was therefore not heuristically labeled as such.

We hope these statistical insights enable researchers to pick specific subsets of scenes for their experiments based on

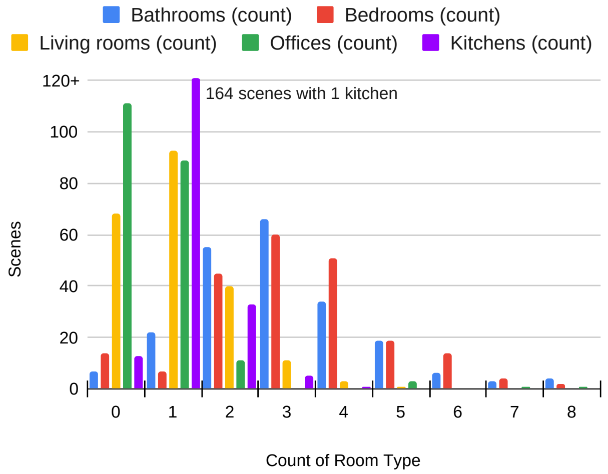

Figure 5. Histogram of room types in each scene based on region instance mapping and category implied room type heuristic. Vertical axis clamped to 120.

relevant criteria. Additionally, further statistical analysis of these data may reveal deeper relationships between objects and their common regions or neighborhoods. The results of this analysis may be useful in downstream tasks such as scene understanding and procedural generation.

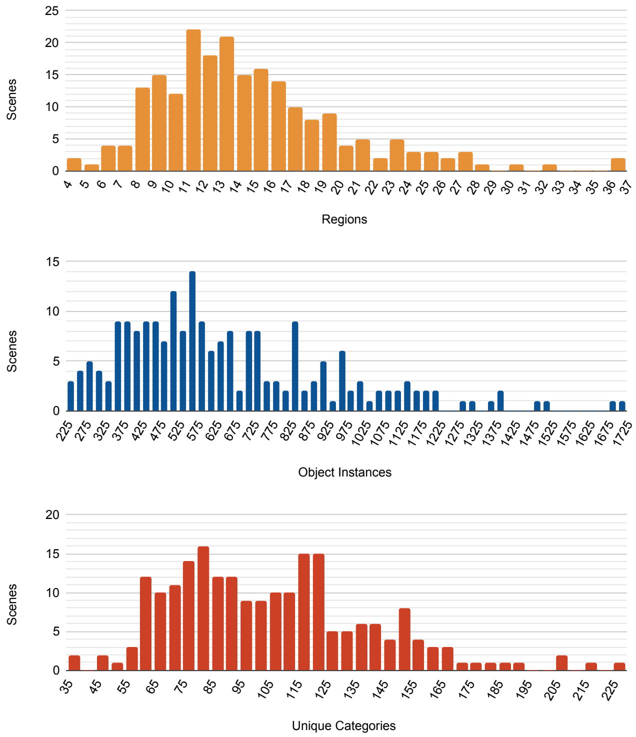

Figure 6. Histograms displaying region, object instance, and unique annotation category counts over all scenes in the dataset.

4. Experiments

In this section, we present experimental results for training Object-Goal Navigation (ObjectNav) policies using the HM3DSEM dataset in the Habitat simulator [2]. To compare the quality of HM3DSEM with prior datasets, we train three different policies (reinforcement, imitation and modular learning) for ObjectNav using three different datasets, HM3DSEM, Gibson [3], and MP3D [6]. We then evaluate each policy on all datasets. For example, a policy trained on HM3DSEM will be evaluated on Gibson and MP3D, even though it was not trained on them. We show that the policies trained on HM3DSEM perform better or are comparable to those trained on Gibson and MP3D when evaluated on all three datasets. This indicates that training ObjectNav policies on HM3DSEM improves cross-dataset domain generalization. We also show that increasing the number of scenes used for training leads to better generalization to previously unseen scenes.

ObjectNav task definition. For our experiments we use an agent matching LoCobot's specification with a base radius of 0.18m and height of 0.88m. The agent is equipped with a 640x480 RGB-D camera (mounted at a height of 0.88m) along with a Compass and a GPS sensor. The agent's action space comprises of the [MOVE_FORWARD, TURN_LEFT, TURN_RIGHT, LOOK_UP, LOOK_DOWN, STOP] actions with a forward step of 0.25m and turn angles of $30°$. We define 6 goal categories (similar to [21]): chair, bed, plant, toilet, tv/monitor, and sofa. The agent is successful if it executes STOP at a location that lies within 1.0m of any object instance from the goal category. While the agent need not directly see the object while stopping, we require that the object can be directly viewed from the stop location without obstruction (i.e., *oracle-visibility* [22]). We evaluate the agent's performance using the standard Success and SPL metrics [1]. Success measures how often the agent finds and stops at the goal object, while SPL measures how efficiently the agent succeeds (i.e., the efficiency of the agent's path relative to the shortest path from the start to goal positions).

ObjectNav episode dataset. We generate episode datasets from the 145 train, 36 val, and 35 test scenes for benchmarking agents on the ObjectNav task. Our episode generation process is similar to prior ObjectNav work [22]. Each episode consists of a scene, a start position where the agent is placed at time $t = 0$, and a goal object category. To generate an episode for a given scene, we uniformly sample a goal from the 6 goal categories. We then randomly sample a start location from the scene that satisfies the following constraints: (1) the start location must be navigable, (2) the goal object must be reachable from the start location, and (3) the distance from the start location to the nearest object from the goal category must lie between 1m and 30m. Following this procedure, we generate episode datasets containing $\sim$ 7.2M train / 1072 val / 1000 test episodes.

	Eval Dataset →	Gibson (val)		MP3D (val)		HM3DSem (val)		Average	
Agent	Train Dataset ↓	Success ↑	SPL ↑	Success ↑	SPL ↑	Success ↑	SPL ↑	Success ↑	SPL ↑
RL	Gibson	25.41	10.22	**18.18**	6.51	28.79	11.32	24.13	9.35
	MP3D	19.66	7.16	17.53	6.59	18.97	6.93	18.70	6.89
	HM3DSem	**32.96**	**13.56**	17.53	**7.48**	**39.36**	**18.02**	**29.95**	**13.03**
IL	Gibson	18.15	7.84	16.57	5.84	20.06	6.03	18.26	6.57
	MP3D	18.95	6.95	**28.57**	**9.91**	25.37	7.73	24.30	8.20
	HM3DSem	**28.02**	**11.64**	25.06	8.72	**33.49**	**11.64**	**28.86**	**10.66**
ML	Gibson	**38.3**	**16.0**	47.1	21.5	53.2	28.5	46.2	22.0
	MP3D	31.9	14.8	45.3	20.5	50.8	21.7	42.67	19.0
	HM3DSem	38.1	15.7	**49.4**	**23.4**	**55.6**	**30.7**	**47.70**	**23.27**

Table 2. **ObjectNav generalization across datasets with different agents.** We report results for a DD-PPO policy (RL), a Behavior Cloning policy (IL), and a modular semantic exploration policy (ML).

				HM3DSem (val)	
	Name	Area (m^2)	Scenes	Success ↑	SPL ↑
HM3DSem	Tiny	3811.76	25	17.79	6.70
	Small	9194.19	80	26.87	11.18
	Medium	12427.08	99	35.16	14.90
	Large	16523.37	145	39.36	18.02

Table 3. **Dataset size vs Performance. (RL)** ObjectNav performance of an RL policy with increasing training dataset sizes.

				HM3DSem (val)	
	Name	Area (m^2)	Scenes	Success (%) ↑	SPL (%) ↑
HM3DSem	Tiny-HD	3811.76	25	34.98	11.84
	Small-HD	5753.60	36	38.90	13.19
	Medium-HD	7099.31	54	48.32	16.07
	Large-HD	9194.19	80	54.20	18.71

Table 4. **Dataset size vs Performance. (IL)** ObjectNav performance of an IL policy with increasing human demonstrations dataset sizes for training.

4.1. Reinforcement Learning

Training Details. We use the architecture proposed in DDPPO [23] to train the ObjectNav agents. A ResNet50 network encodes the RGB-D images into visual representations, which are concatenated with the embeddings of the goal object category, the previous action, and the GPS+Compass sensor readings. This joint embedding is passed to a 2-layer 512-D LSTM network. The output of the LSTM module is passed to fully-connected layers, which predict the action probabilities and state values. The agent is trained for 400M steps, after which it overfits on the training dataset.

Scene Dataset Comparison. The top 3 rows in Table 2 show the performance of DDPPO agents trained on the Gibson, MP3D, and HM3DSem episodes. We evaluate these agents on the validation scenes from all three datasets. We observe that training on HM3DSem leads to the best performance averaged over all datasets. Surprisingly, the agent trained on HM3DSem outperforms the agent trained on Gibson when evaluated on Gibson. This indicates that improved visual reconstruction and annotation accuracy in HM3DSem leads to policies that generalize better even across datasets.

Dataset Size vs Performance. We also train agents on different subsets of HM3DSem scenes to study the effects of dataset scaling. In Table 3, we observe that the performance of our agent improves with more training scenes, with HM3DSem-Large performing more than twice as good as HM3DSem-Tiny (39.36% vs 17.70%).

4.2. Imitation Learning

Training Details. We collect $77k$ human demonstrations for 80 HM3DSem training scenes using Habitat-Web [24]. Following [24] we use a simple CNN+RNN architecture. For RGB, we use a ResNet18 [25] that is randomly initialized. For depth, we use a ResNet50 which was pre-trained on PointGoal navigation task using DD-PPO [23] on gibson dataset. The GPS+Compass inputs are passed through fully-connected layers to embed them to 32-D vectors. In addition to RGB-D and GPS+Compass, we use two additional semantic features following [26] – semantic segmentation of the input RGB observation, predicted using a RedNet [27] model, and a 'Semantic Goal Exists' feature which is the total area of the visual input occupied by the goal object category. To predict the semantic features, we use the RedNet model from [26], which was pre-trained on SUN RGB-D [28] and finetuned on $100k$ randomly sampled views from MP3D scenes. Finally, we also feed in the object goal category embedded into a 32-D vector. These input features are concatenated to form an observation embedding, and fed into a 2-layer, 2048-D GRU at every timestep. We train this policy for $\sim$400M steps, which amounts to $\sim$20 epochs on $\sim 77k$ demonstration episodes.

Scene Dataset Comparison. Table 2 (rows 4-6) shows the performance of ObjectNav policies trained using imitation learning (specifically, behavior cloning) on $10k$ human demonstrations from the HM3DSem, Gibson and MP3D datasets. We evaluate these agents on the validation scenes from all three datasets. We find that the imitation learning policy trained on HM3DSem achieves the best performance averaged across all validation datasets. In fact, the HM3DSem policy even outperforms the Gibson policy when evaluated on Gibson validation scenes. This echoes our findings from Sec. 4.1 and further emphasizes the high annotation quality in HM3DSem.

Dataset Size vs Performance. In Table 4, we study the dataset scaling behavior by training on different subsets of HM3DSem scenes, ranging from 25 to 80 scenes. We observe consistent improvements in the validation performance as we increase the number of training scenes. This suggests that it is valuable to collect large-scale human demonstrations for ObjectNav and that the performance is likely to improve further as we increase the number of training scenes.

4.3. Modular Learning

In addition to end-to-end reinforcement and imitation learning, modular learning has emerged as a popular alternative for training policies to tackle various Embodied AI tasks [21, 29–38]. Besides showing that training on HM3DSem leads to better end-to-end navigation policies, we also show that it leads to better modular components. Specifically, we train the Goal-Oriented Semantic Exploration (SemExp) policy of [21] on HM3DSem, Gibson, and MP3D and evaluate its generalization to other datasets.

Training Details. The approach proposed in [21] builds a top-down semantic map by projecting first-person semantic segmentation predictions with depth, selects an exploration goal as a function of the semantic map and the goal object with a learned exploration policy, and plans low-level actions to reach this goal. We replicate the exploration policy architecture and training process of [21] for all datasets. We use Mask-RCNN [21] pre-trained on MS-COCO for object detection and instance segmentation. The semantic map has a shape K x M x M matrix where M x $M =$ 960 x 960 is the map size, with each cell corresponding to 25 cm^2 (5 cm x 5 cm) in the physical world, and $K = 16$ is the number of map channels. Semantic map features are computed with a convolutional neural network and passed through a feed-forward neural network along with a learnable embedding for the goal object to compute an exploration goal in $[0, 1]^2$, which is then converted to top-down map space. The exploration policy is trained for 10 million steps using reinforcement learning with the Proximal Policy Optimization algorithm [39], and the distance reduced to the nearest goal object as the reward. As in [21], we sample the long-term goal at a coarse time scale once every 25 steps.

Scene Dataset Comparison. Table 2 (bottom three rows) shows the performance of agents with a semantic exploration policy trained on HM3DSem, Gibson, or MP3D training scenes and evaluated on each dataset's validation scenes. Agents trained on HM3DSem scenes achieve the best validation performance averaged across all datasets.

4.4. ObjectNav Challenge 2022

The Habitat ObjectNav challenge [40] in the Embodied AI workshop in CVPR 2022 used the HM3DSem dataset. The challenge received a total of 1022 submissions from 54 teams through the course of the challenge. This is much higher in comparison to the 2021 and 2020 Habitat ObjectNav challenge which received a total of 400 and 563 submissions from 45 and 27 teams respectively. The task definition was identical between the 2020-21 and 2022 challenges, and the only change was the dataset from Matterport3D [6] to HM3DSEM. The increase in participation highlights the importance of improving dataset scale and quality for community adoption. We report the final Success Rate and SPL of the submission from the top 5 teams and the DDPPO baseline on the Test-Challenge split in Table 5.

Team Name	**Success Rate** (%) ↑	**SPL** (%) ↑
ByteBOT	64.0	35.0
BadSeed	65.0	33.0
elf	61.0	30.0
Populus A.	60.0	30.0
Stretch	56.0	29.0
DDPPO	25.0	12.0

Table 5. **Habitat 2022 ObjectNav Challenge.** Performance of the top entries to the Habitat 2022 ObjectNav Challenge on the Test Challenge split. Entries are ordered by SPL.

5. Conclusion

We present the HM3DSem dataset which is the largest public dataset of real-world spaces with dense semantic annotations. Unlike prior datasets, HM3DSem uses texture information to annotate pixel-accurate object boundaries. The dataset has undergone an intense expert annotation as well as a verification process to maximize accuracy and coverage. All scene annotations are provided in a standardized format, making it easy to use with the existing Habitat simulator. We demonstrate the effectiveness of the HM3DSem dataset for the Object Goal Navigation task using reinforcement learning, imitation learning, and modular learning methods. Across different methods, we show the importance of improved annotation quality and larger datasets by showing that policies trained using HM3DSem outperform the policies trained on prior datasets and the performance of policies improves as we increase the training dataset size. The introduction of the HM3DSem in the Habitat ObjectNav challenge in 2022 led to a significant increase in participation. We hope that the high quality and scale of HM3DSem spurs future progress in Embodied AI.

Acknowledgements

The Georgia Tech effort was supported in part by NSF, ONR YIP, and ARO PECASE. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

References

[1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 1, 6

[2] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 6

[3] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martın-Martın, Linxi Fan, Guanzhi Wang, Shyamal Buch, Claudia D'Arpino, Sanjana Srivastava, Lyne P Tchapmi, Kent Vainio, Li Fei-Fei, and Silvio Savarese. iGibson, a simulation environment for interactive tasks in large realistic scenes. *arXiv preprint*, 2020. 3, 6

[4] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *arXiv preprint arXiv:2106.14405*, 2021. 2, 3

[5] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=-v4OuqNs5P. 1

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *Fifth International Conference on 3D Vision (3DV)*, 2017. 1, 2, 3, 4, 5, 6, 8

[7] Afshin Dehghan, Gilad Baruch, Zhuoyuan Chen, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, and Elad Shulman. ARKitScenes-a diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data, 2021. URL https://openreview.net/pdf?id=tjZjv_qh_CE. 1, 2, 3

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1, 3

[9] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. RIO: 3D object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019. 2, 3

[10] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5664–5673, 2019. 2, 3

[11] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 3

[12] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-Thor: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017. 2

[13] Claudia Yan, Dipendra Misra, Andrew Bennnett, Aaron Walsman, Yonatan Bisk, and Yoav Artzi. Chalet: Cornell house agent learning environment. *arXiv preprint arXiv:1801.07357*, 2018. 2

[14] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. VirtualHome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018. 2

[15] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1754, 2017. 2

[16] Huan Fu, Bowen Cai, Lin Gao, Lingxiao Zhang, Cao Li, Zengqi Xun, Chengyue Sun, Yiyun Fei, Yu Zheng, Ying Li, et al. 3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics. *arXiv preprint arXiv:2011.09127*, 2020. 2

[17] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X. Chang, and Matthias Nießner. Scan2CAD: Learning CAD model alignment in RGB-D scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[18] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. RoboTHOR: An open simulation-to-real embodied AI platform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3164–3174, 2020.

[19] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, Sai Bi, Zexiang Xu, Hong-Xing Yu, Kalyan Sunkavalli, Milos Hasan, Ravi Ramamoorthi, and Manmohan Chandraker. OpenRooms: An end-to-end open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[20] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 2

[21] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020. 6, 8,

11

[22] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020. 6

[23] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. DD-PPO: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *International Conference on Learning Representations (ICLR)*, 2020. 7

[24] Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[26] Joel Ye, Dhruv Batra, Abhishek Das, and Erik Wijmans. Auxiliary Tasks and Exploration Enable ObjectNav. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2021. 7

[27] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. RedNet: Residual Encoder-Decoder Network for indoor RGB-D Semantic Segmentation. *arXiv preprint arXiv:1806.01054*, 2018. 7

[28] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun RGB-D: A RGB-D scene understanding benchmark suite. In *Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 7

[29] Devendra Singh Chaplot, Saurabh Gupta, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural mapping. *8th International Conference on Learning Representations, ICLR 2020*, 2020. 8

[30] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological slam for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12875–12884, 2020.

[31] Devendra Singh Chaplot, Helen Jiang, Saurabh Gupta, and Abhinav Gupta. Semantic curiosity for active visual learning. In *ECCV*, 2020.

[32] Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, and Oleksandr Maksymets. Waypoint models for instruction-guided navigation in continuous environments. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[33] Devendra Singh Chaplot, Murtaza Dalal, Saurabh Gupta, Jitendra Malik, and Ruslan Salakhutdinov. Seal: Self-supervised embodied active learning. In *NeurIPS*, 2021.

[34] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation. In *ICLR*, 2022.

[35] Meera Hahn, Devendra Chaplot, Shubham Tulsiani, Mustafa Mukadam, James Rehg, and Abhinav Gupta. No rl, no simulation: Learning to navigate without navigating. 2021.

[36] So Yeon Min, Devendra Singh Chaplot, Pradeep Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. Film: Following instructions in language with modular methods. In *International Conference on Learning Representations (ICLR)*, 2022.

[37] Gabriel Sarch, Zhaoyuan Fang, Adam W. Harley, Paul Schydlo, Michael J. Tarr, Saurabh Gupta, and Katerina Fragkiadaki. Tidee: Tidying up novel rooms using visuo-semantic commonsense priors. In *European Conference on Computer Vision (ECCV)*, 2022.

[38] Santhosh K. Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Computer Vision and Pattern Recognition (CVPR), 2022 IEEE Conference on*. IEEE, 2022. 8

[39] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017. 8

[40] Karmesh Yadav, Santhosh Kumar Ramakrishnan, John Turner, Aaron Gokaslan, Oleksandr Maksymets, Rishabh Jain, Ram Ramrakhya, Angel X Chang, Alexander Clegg, Manolis Savva, Eric Undersander, Devendra Singh Chaplot, and Dhruv Batra. Habitat challenge 2022. `https://aihabitat.org/challenge/2022/`, 2022. 8