Pre-Trained Masked Image Model for Mobile Robot Navigation

Vishnu Dutt Sharma, Anukriti Singh, Pratap Tokekar

***Abstract*— 2D top-down maps are commonly used for the navigation and exploration of mobile robots through unknown areas. Typically, the robot builds the navigation maps incrementally from local observations using onboard sensors. Recent works have shown that predicting the structural patterns in the environment through learning-based approaches can greatly enhance task efficiency. While many such works build task-specific networks using limited datasets, we show that the existing foundational vision networks can accomplish the same without any fine-tuning. Specifically, we use Masked Autoencoders, pre-trained on street images, to present novel applications for field-of-view expansion, single-agent topological exploration, and multi-agent exploration for indoor mapping, across different input modalities. Our work motivates the use of foundational vision models for generalized structure prediction-driven applications, especially in the dearth of training data. We share more qualitative results at https://raaslab.org/projects/MIM4Robots.**

I. INTRODUCTION

Mobile robot navigation through unknown areas has been studied by the robotics community for a long time [27]. Generally, in the existing approaches, the robot updates the map based on its observations so far and moves according to the task at hand, such as PointGoal navigation, ObjectGoal navigation, and exploration [2]. In the case of ground robots, the map is typically represented in a top-down view or Bird's Eye View (BEV), as the robot motion is constrained on the ground plane. Aerial robots also use Top-down representations for navigation and exploration or to help others when working in tandem with other aerial or ground robots [1], [20], [36], [37].

The traditional approach is to build a map by fusing the robot's observations. Recent works across the wider robotics community have started exploring learning-based approaches to augment the robot's onboard data about the environment, e.g., occupancy map, point cloud, etc., to accomplish tasks [11], [21], [22], [33], [35]. These methods learn the patterns in representations and can predict the yet-unobserved regions based on partially observed environments. The prediction can then be used to make informed decisions for safer and more efficient motion planning.

Learning-based methods require extensive datasets, which are challenging to get in robotics applications compared to computer vision. Simulators can generate virtual data, but face a sim2real gap. Many computer vision methods trained on large datasets may not directly apply to robot applications due to distributional differences in image representation; computer vision datasets are mainly comprised of first-person views, captured from a height often taller than the camera mounted on ground robots. Fine-tuning often seems to be the solution but requires similarity between pre-training and fine-tuning tasks, which can be challenging for robot navigation representations such as top-down images, semantic maps, and occupancy maps.

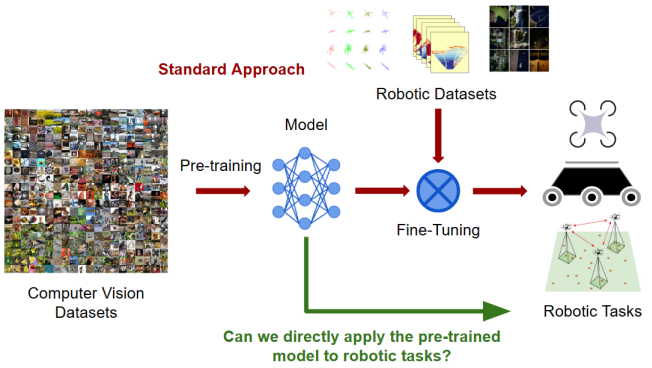

Fig. 1: Traditional approach of leveraging models trained on huge computer vision datasets can be applied to robotic tasks reliant on top-down images, albeit with some task-specific fine-tuning. We show that this is not necessary and some models, such as MAE [19] can be applied directly to these robotics tasks.

The recent emergence of self-supervised foundational models, which are trained on huge datasets, aims to achieve generalizability by leveraging a diverse distribution of datasets. This approach is premised on the belief that it should prompt the model to reason about fundamental concepts such as shapes and textures. However, the datasets used may not necessarily include the same distribution of images that we expect to observe during robot navigation.

This raises the question: *Can we apply pre-trained computer vision models directly on robotics tasks such as navigation and exploration without fine-tuning?* Surprisingly, the answer is yes. We substantiate this assertion with a masked image model that learns to reconstruct an image using representation learning. Specifically, Masked Autoencoder (MAE) [19], which randomly patches the image to learn local correlation and reconstruct the masked parts. We show how, despite being trained on first-person view [9] images, it can make reasonable predictions about the unseen areas in *top-down* RGB, semantic, and occupancy maps, which improves 2D planning for efficient robot navigation. We find that there is no need to fine-tune MAE on specific tasks

This work is supported by the National Science Foundation under Grant No. 1943368 and ONR under grant number N00014-18-1-2829.

The authors are with the Dept. of Computer Science, University of Maryland, College Park, MD, USA {vishnuds, anukriti, tokekar}@umd.edu

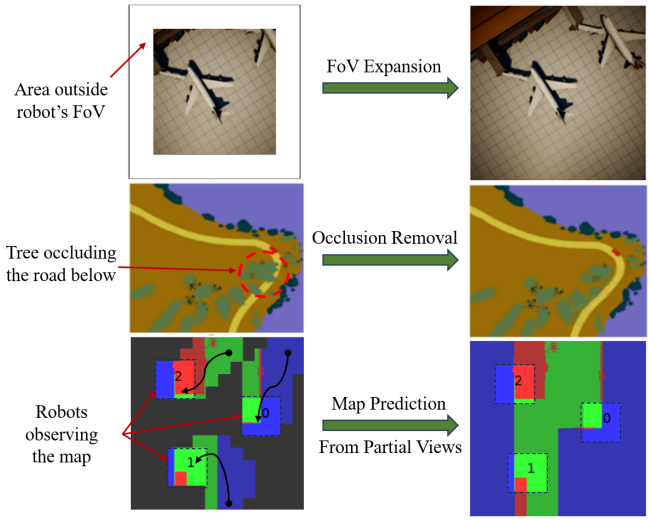

Fig. 2: Example of robotics tasks solved with help of Masked Autoencoder.

for improvement, making it further appealing for robotics applications that may not have adequate training data.

Specifically, we make the following contributions in this paper:

- We study MAE as an expainting network for top-down images across RGB, semantic maps, and binary maps, and present quantitative and qualitative results across various degrees of increasing field-of-view for indoor and outdoor images.
- We present a novel uncertainty-driven exploration method for 2D semantic map reconstruction using MAE and compare it to non-predictive approaches to highlight the benefits of structural pattern prediction.
- We show that MAE can be effectively applied for a case study of single robot navigation aided by occupancy prediction, resulting in more efficient operation compared to a standard, non-predictive baseline method.

Our work highlights how foundational self-supervised learning algorithms like masked image model (MAE) can be used for robot tasks by choosing appropriate modalities without any fine-tuning, and paves the way for further improvement to the existing capabilities by task-specific tuning of these models. Coupled with its applicability to a variety of robotics applications, as shown in Fig. 2, MAE could potentially be the free-lunch all-around solution for 2D map-based navigation.

II. Related Works

A. Mapping for Robot Navigation

Top-down images and map representations are vital for robot navigation and exploration. Navigating through an unknown map by Simultaneous Mapping and Localization (SLAM), which utilizes the robot's past observations, has been a cornerstone of robotics for robotics. A top-down semantic map is another representation of interest for robotic applications. These maps are useful for semantic goal navigation [17], [18]. Top-down images are also beneficial for aerial robot tasks such as surveying and scouting [8], [28]. The maps obtained by the aerial robots can be used to help the ground robots navigate. Semantic maps are obtained from such images to identify navigable and non-navigable areas for the ground robot.

Recent works in this domain have sought to improve task efficiency by *predicting* the unobserved regions of the map to plan ahead. 2D Occupancy map, a top-down representation, has been the focus of many of these works, showing improvement in navigation, exploration distance, and time [22], [29], [34], [40]. Katyal et al. [23] show these benefits for high-speed navigation, highlighting the importance of predictions. While the predictions are limited to the perception module, it can also enhance planning by extracting uncertainty from the predictions [16], [21]. The idea of uncertainty extraction also proves helpful in heterogeneous robot teams for risk-aware planning [36]. The key challenge with all these systems is that they need to be trained on the appropriate modalities, for which sufficient data may not be available, leading us to ask if there exist pre-trained models that can be used in these applications without much training effort, or better, without any fine-tuning at all?

B. Self-supervised masked encoding

In recent times, various approaches like BEiT [4], iBOT [41], and ADIOS [38] have drawn inspiration from masked language models. These methods have demonstrated remarkable competitiveness in the realm of self-supervised learning (SSL). All three techniques leverage vision transformers and propose strategies to "inpaint" images that have been partially obscured by random masks in various ways. The idea of map prediction is similar to this, and existing works for robotic applications rely on generative models [25], [30], [31], which require training or fine-tuning networks on simulation data to get accurate results.

Masked Autoencoder (MAE) uses Vision Transformer (ViT) encoder [12] and is trained to use only the visible patches of an image to predict the missing patches, similar to the training strategy of BERT [10]. MAE uses linear projections and position encodings for feature representation and is trained with mean squared error (MSE) between the reconstructed and original images in the pixel space, but only for masked patches. While MAE is also trained on RGB images only from the ImageNet-1K dataset [9], the underlying ViT architecture allows it to reason about other modalities as shown by MultiMAE [3]. Therefore, we use MAE for our study and show its effectiveness for prediction and inpainting across various modalities in top-down images useful for robotic tasks, without any finetuning.

III. Methodology

We aim to determine whether the pre-trained masked autoencoder can effectively predict unobserved regions on 2D maps, represented as top-down RGB images, semantic maps, or binary maps. We focus on three tasks relevant to

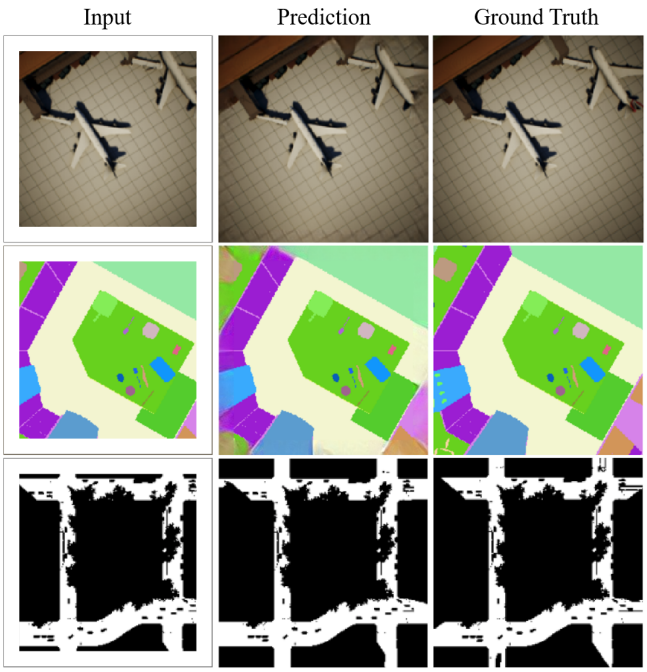

Fig. 3: Masked Autoencoder can be used to expand the effective FoV in top-down RGB, semantic, and binary images without fine-tuning.

Fig. 4: Results of expanding FoV for indoor images in three masking scenarios. The corner of the bathtub and room is accurately predicted based on the symmetry of the lines.

robot navigation and exploration, with detailed descriptions provided in the following subsections.

A. FoV Expansion and Navigation

Katyal et al. [22] conducted a comprehensive study on various convolutional networks to augment the effective Field of View (FoV) of the robot for predicting unexplored occupancy maps in the robot's surroundings. In their subsequent work, they demonstrated that the prediction of future occupancy maps can improve high-speed navigation [21]. This research employed U-Net [32] as an image-to-image translation network for occupancy map prediction, founding the basis of subsequent research to further enhance robot navigation and exploration [16], [29], [35], [40].

In this study, we primarily investigate the FoV expansion task, as shown in Fig. 3 and 4. Instead of employing raw occupancy maps, we opt for semantic segmentation maps and binary maps, modalities that are eventually used by conventional robotic planners. Additionally, we study RGB images, a modality consistent with the one used for MAE training and relevant to aerial mapping and surveying applications. This allows us to examine (a) whether MAE can work well on a different camera view, and (b) how other modalities, i.e., semantic and binary maps, perform during inference when compared with the one used for training MAE. Furthermore, we extend the original study by evaluating MAE performance in both indoor and outdoor environments. The inputs to MAE are provided as 3-channel images, with labels replaced by corresponding colors in semantic and binary maps. Subsequently, the colors in the output images are reconverted to labels by substituting them with the label associated with the closest color in the input images. These labels are then utilized as the assigned classes for evaluation.

B. Multi-Agent Uncertainty Guided Exploration

Uncertainty-guided navigation and exploration, as proposed in previous works [16], [21], [36] aims to enhance active robot exploration by combining the uncertainty-driven exploration technique with image inpainting networks. The eventual goal is to efficiently map the whole environment. These tasks, however, limit themselves to single-agent applications. We propose a novel approach along these lines for a multi-agent setup with pre-trained MAE, without any architectural changes such as dropout injection [15].

Our approach draws on concepts from bootstrapping [13] and adversarial attacks on neural networks [39]. By injecting minimal random noise into the input image, we obtain predictions on perturbed inputs from MAE. Despite resulting in imperceptible visual changes. We repeat this procedure multiple times to get n predictions on such *bootstrapped* inputs from MAE and find variance across each pixel, summed over the channels, as the uncertainty in prediction. Conceptually, pixels with high variance indicate regions where MAE lacks strong structural cues from visible input, necessitating direct observations from the robot.

We put this premise to the test by looking at the prediction accuracy at each step of the exploration. To execute exploration, unexplored locations and those with high uncertainty are subsequently grouped together to identify distinct regions for potential exploration. The robots are assigned to

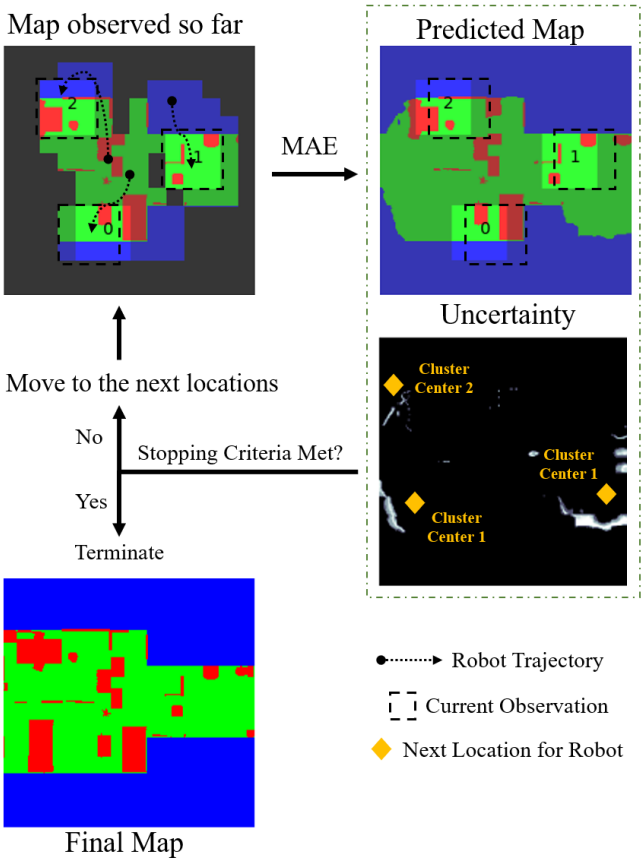

Fig. 5: An overview of the multi-agent exploration pipeline.

this cluster based on their proximity to the cluster center while ensuring that no two robots are assigned to the same region. We stop the exploration when the cluster centers stabilize. Fig. 5 shows an overview of this process. This study addresses two critical questions: (a) how to extract uncertainty from MAE, a point-prediction network, and (b) can predictions be leveraged to fill gaps in unexplored maps when resource constraints, such as battery limitations for aerial robots, hinder complete coverage?

We compared the following algorithms for this task:

- **Boustrophedon Cell Decomposition Algorithm (Lawnmower)**: Proposed by Choset et al. [6], this algorithm divides the regions into n contiguous scanlines of similar size, each assigned to one robot. Each robot scans the designated area for coverage. For this method, we position the robots at the start of the respective scanlines to streamline the process.
- **KMeans Clustering (KMeans-U)**: Here, we apply the KMeans algorithm (with n centers) to the Cartesian coordinates of the unexplored grid cells to identify the centers of the unexplored regions. The robots are then assigned to these regions based on proximity to the cluster and move towards them. At each step, the robots observe the region below and include it in the known map. Then we repeat the clustering process to find centers for the remaining unexplored areas.
- **KMeans Clustering followed by Reconnaissance (KMeans-R)**: Employing KMeans directly may lead to unexplored regions at the center of the map when the cluster centers stabilize (which is a stopping criterion). To address this, we introduce an additional step of relocating all the robots to the center of the map after stabilization. This results in enhanced coverage at the expense of time.
- **KMeans Clustering on Unknown and Uncertain Regions (KMeans-U^2)**: In this method, we extract uncertainty from MAE and use the locations with non-zero variance, along with those yet unexplored, for clustering. The procedure for assigning clusters to robots follows a similar approach as in the earlier methods.

Here we aim to assess the prediction capabilities of MAE and thus make predictions on the map explored so far at each step. We compare the distributions of coverage to reach 95% prediction accuracy to find which algorithm is more efficient in predicting the unexplored map.

C. Navigation with Prediction

For autonomous navigation, it is crucial to know the map of the environment. The classical methods treat the unexplored area as unknown and build a costmap on the basis of only the current observation. The robot can traverse to the edge of the frontier before needing another observation to plan the path ahead. Effectively, the sensor range of the robot defines the maximum distance it can traverse at once. Previous works have shown that predicting future occupancy can result in faster navigation [24] and smoother control [14]. These works train neural networks to make these predictions, using synthetically generated data and real-world data obtained by running a robot around. While the former may run into Sim2Real issues, the data collection with the latter is an arduous process. We test if pre-trained MAE could instead be used for prediction while side-stepping the data issues.

For this, we use the predictions of MAE with a standard path planning algorithm on multiple indoor floor plans. The robot starts from its initial position where the rest of the area in the map is hidden; using MAE we reconstruct the unseen map and update the path at every next step. The predicted unseen map acts as an estimate of the occupancy ahead, which helps to reconstruct an informed costmap for navigation. This helps the robot cover a greater distance at once by moving to the frontier of every step. Figure 6 shows one such example where the robot is able to plan a shorter path ahead since the predictions help to know the shape of the obstacle before the robot actually explores that area.

We compare our prediction-based approach with a non-predictive approach and calculate the number of steps (and observations) needed to reach a pre-defined goal.

IV. Experimental Setups and Evaluation

In this section, we describe the experimental setup and our findings for each task defined in Section III. Throughout our experiments, we utilize the MAE based on ViT-Large trained on ImageNet-1K dataset [9].

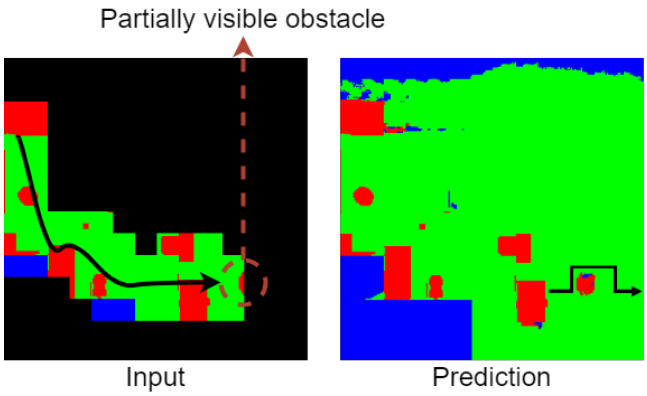

Fig. 6: Left: The area robot has explored till now. Right: Prediction of obstacle (red) shape aiding robot path planning.

A. FoV Expansion

To study FoV expansion, we mask the periphery of the given image by different amounts. MAE uses patches of size 16×16 pixels, and masking a patch requires all the pixels in the patch to be masked. We mask the images with 1-3 patches on each side, resulting in an expansion of **1.17x**, **1.4x**, and **1.75x** to the robot's perceptual range, i.e., the number of pixels in a direction if the robot is at the center. As the data used by Katyal et al. [22] is not publicly available, we perform an evaluation on the dataset collected from two photorealistic simulated environments, consisting of diverse indoor and outdoor scenes.

Indoor Data: For indoor environment, we use AI2-THOR [26] which has 120 indoor scenes such as kitchens, living rooms, bathrooms, etc. We collect 1444 RGB and segmentation images with a top-down camera of a field of view of 80 degrees and rotated at intervals of 30 degrees (e.g., 30, 60, 90, etc.).

Outdoor Data: For outdoor images were taken from AirSim VALID dataset [5] which consist of scenes from cities, suburbs, and mountains among others, captured at different altitudes from an aerial robot. We sample 1000 images from this dataset for this study. For these environments, we also evaluate MAE on binary images, consisting of navigable and non-navigable regions, as a stand-in for occupancy maps.

We evaluate the RGB predictions for the FoV increase on the following metrics typically used to quantify visual similarity: (1) Frechet Inception Distance (FID), (2) Structural Similarity Index Measure (SSIM), (3) Peak Signal-to-Noise Ratio (PSNR), and (4) Mean Squared Error (MSE). For the semantic and binary images, we use mean Intersection-over-Union (mIoU) as the key metric but also provide the results for some of the aforementioned metrics since we use MAE to predict visually similar images for these modalities.

Results: Table I summarizes the results for RGB images for both types of environments. We find that increasing the FoV results in worse results than expected since MAE, an inpainting network can not reliably predict the outside areas without much context. 1.75x expansion is the extreme case where the predictions get blurry. Figure 3 and Figure 4 show some examples in RGB outdoor and indoor scenes respectively and highlight this effect.

Table II and Table III summarize results for semantic and binary maps. The mIoU is very high for 1.17x expansion and goes down with increasing FoV. The effect is worse indoors as it contains many more classes (270) compared to outdoors (30) and thus may not reliably perform color-to-label matching. Also, small objects are within the scene and on the periphery, and MAE can not expand them without seeing some part of them. Note that the mIoU here is not weighted by the labels' population size. Predictions on binary maps are relatively more robust since the size of objects in each class and the difference in color mapping are larger than the semantic maps. These results present an encouraging picture for a network that was not trained on such images. We note that Katyal et al. [22] report a maximum SSIM of 0.523, 0.534, and 0.504 on real-world data for similar expansion factors. MAE results in better SSIM on both semantic segmentation and binary maps in comparison. Katyal et al. [22] report higher numbers, 0.899, 0.0818, and 0.760, on synthetic data which is similar to the distribution used for training their network. We find MAE on semantic segmentation maps still achieves higher SSIM. However, MAE with binary maps do not achieve similar performance, but still produce good results despite being trained on a different modality and camera view.

TABLE I: Results for increasing the FoV in RGB images

Setup	Expansion	FID ↓	SSIM ↑	PSNR ↑	MSE ↓
Indoor	1.17x	**17.83**	**0.94**	**27.76**	**13.76**
	1.40x	41.79	0.86	22.23	32.42
	1.75x	76.59	0.78	19.18	52.98
Outdoor	1.17x	**53.66**	**0.84**	**26.38**	**33.59**
	1.40x	77.91	0.69	22.79	49.91
	1.75x	116.09	0.55	19.98	67.80

TABLE II: Results for increasing the FoV in Semantic segmentation images

Setup	Expansion	mIoU ↑	FID ↓	SSIM ↑	PSNR ↑
Indoor	1.17x	**0.86**	**43.48**	**0.94**	**23.06**
	1.40x	0.55	75.42	0.84	17.33
	1.75x	0.34	110.01	0.78	14.90
Outdoor	1.17x	**0.90**	**42.63**	**0.94**	**25.96**
	1.40x	0.73	73.03	0.86	21.39
	1.75x	0.57	118.56	0.79	18.80

TABLE III: Results for increasing the FoV in Binary images from Outdoor environment

Expansion	mIoU ↑	FID ↓	SSIM ↑	PSNR ↑
1.17x	**0.90**	**51.87**	**0.95**	**30.36**
1.40x	0.78	88.44	0.76	22.05
1.75x	0.64	120.94	0.56	17.81

B. Multi-Agent Uncertainty Guided Exploration

For this task, we use 3-channel semantic map representations, consisting of free, occupied, and out-of-boundary

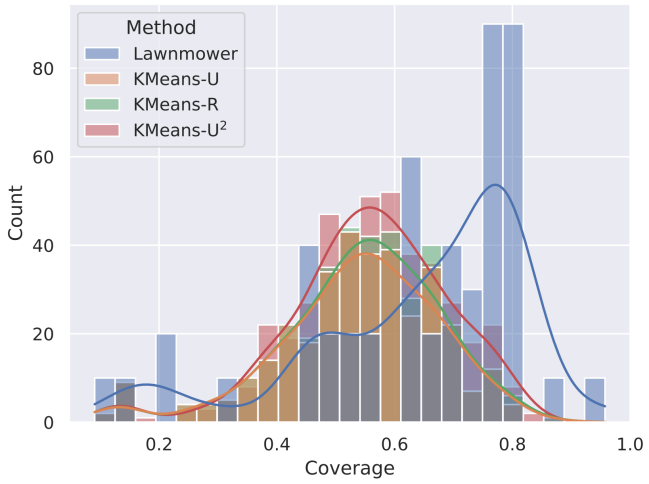

Fig. 7: Comparison between the multi-agent exploration algorithms to reach at least 95% accuracy in prediction of the unexplored map.

regions, using color-to-label matching on the MAE prediction for labeling, as described in Section III-A. In our experiments, we utilize 50 living room scenes from the ProcTHOR [7] framework. We convert their ground truth semantic segmentation maps for the 3-class labeling. These labeled images are transformed into 3-channel RGB images, with free, occupied, and out-of-boundary regions represented by green, red, and blue colors, respectively.

We use $n = 3$ aerial robots and conduct 10 experiments in each room, resulting in 500 runs total for each algorithm. We select the initial positions of the robots randomly. We assume that the area of the room to be explored is known beforehand and that the robots fly at a height taller than the obstacles and do not collide with each other. Each map is represented as an image with dimensions 224×224 pixels, and each robot can observe an area of size 48×48 pixels around it. We treat this as a centralized task, and the observations from all the robots are combined for decision-making.

Results: To compare the methods, we look at the distribution of coverage to reach at least 95% accuracy in predicting the whole map given the partial observation. We visualize our findings in Fig. 7. As shown, most runs with Lawnmower need to cover around 75-85% of the area. This happens due to the naive movement pattern of the robots with Lawnmower and thus the robots do not benefit from the inpainting capability of MAE. All KMeans algorithms, on the other hand, are able to take advantage of it and therefore most runs with them need only 50-60% coverage to reach the same accuracy. KMeans-U^2 is especially denser here as it guides the robots to areas with uncertainty, reducing the chances of incorrect predictions. We note that some heavy-tailed behavior is observed in these plots as some rooms are very simple, and a few predictions may be enough to make good predictions in them. Additionally, spawning robots at the start of the scanlines with Lawnmower places them far apart initially, an advantage other algorithms do not enjoy. This results in Lawnmower sometimes getting better accuracy with less coverage in a simple environment

These findings highlight an intriguing observation about regions with regularly shaped objects: most shapes can be reasonably inferred by looking only at a part of them. As a result, areas with such objects may not be as beneficial for exploration after partial observation, as the large unexplored regions are. KMeans-U^2 performs better as it prioritizes exploring unexplored regions only when it can make a confident (low variance) estimate about objects based on the partial view. The effectiveness of this approach hinges on a prediction model's accuracy in making precise predictions, a task which our experiments have shown MAE excels at.

C. Navigation with prediction

For this application, we use a setup similar to the previous task. Specifically, we represent the occupancy map as a 3-channel semantic map and use color-to-label matching on the MAE predictions. We select 5 large living room scenes from ProcTHOR [7] and choose 20 start-goal pairs on them, located far apart. The map size, robot's field of view, and prediction input are similar to Section IV-B. The unseen area is predicted by MAE. Using the predicted segmented map, we generate a costmap and use A* path planning algorithm for navigation. This process is repeated till the robot reaches the goal.

Results: Across twenty generated paths, our method takes on an average 10.5 steps with a standard deviation of 2.9 steps, whereas the traditional method takes 21.6 steps with a standard deviation of 7.3 steps. It is worth noting that with predictions, the larger frontier helps the robot estimate the shape of an obstacle beforehand, based on partial views, which leads to a reduction of 48% in the total number resulting in efficient navigation.

V. Limitations and Future Work

In this work we show how MAE, a self-supervised network, pre-trained on first-person-view images can be applied to prediction-augmented robotic tasks reliant on top-down maps, without any fine-tuning. Our experiments show its applicability across various robotic tasks, involving different types of input modalities. A key takeaway from our work is that such models are capable of reasoning about regular geometric shapes and directly benefit robots in an environment filled with such patterns. We hope our analysis paves the way for further studies and the development of applications based on such powerful models.

Our work focuses on the efficacy of a pre-trained model and is especially suitable for applications suffering from a lack of training data. We expect improvement in results with task-specific fine-tuning in future works. A drawback of MAE is that it requires the mask to be composed of square patches. While this can be attained with some innovative engineering when the required masks are irregular, other models supporting unrestricted masking might be more suited for this job. Whether they are able to retain the benefits of MAE or not will be explored in our future work.

References

[1] Dario Albani, Daniele Nardi, and Vito Trianni. Field coverage and weed mapping by uav swarms. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4319–4325. Ieee, 2017.

[2] 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.

[3] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 348–367. Springer, 2022.

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[5] Lyujie Chen, Feng Liu, Yan Zhao, Wufan Wang, Xiaming Yuan, and Jihong Zhu. Valid: A comprehensive virtual aerial image dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2009–2016, 2020.

[6] Howie Choset and Philippe Pignon. Coverage path planning: The boustrophedon cellular decomposition. In *Field and service robotics*, pages 203–209. Springer, 1998.

[7] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.

[8] Jaime del Cerro, Christyan Cruz Ulloa, Antonio Barrientos, and Jorge de León Rivas. Unmanned aerial vehicles in agriculture: A survey. *Agronomy*, 11(2):203, 2021.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Harnaik Dhami, Vishnu Dutt Sharma, and Pratap Tokekar. Pred-nbv: Prediction-guided next-best-view planning for 3d object reconstruction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, (in press).

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

[14] Amine Elhafsi, Boris Ivanovic, Lucas Janson, and Marco Pavone. Map-predictive motion planning in unknown environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8552–8558. IEEE, 2020.

[15] Y Gal and Z Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning [eb/ol]. *arXiv preprint arxiv:1506.02142*, 2015.

[16] Georgios Georgakis, Bernadette Bucher, Anton Arapin, Karl Schmeckpeper, Nikolai Matni, and Kostas Daniilidis. Uncertainty-driven planner for exploration and navigation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 11295–11302. IEEE, 2022.

[17] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation. *arXiv preprint arXiv:2106.15648*, 2021.

[18] Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15460–15470, 2022.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[20] Shannon Hood, Kelly Benson, Patrick Hamod, Daniel Madison, Jason M O'Kane, and Ioannis Rekleitis. Bird's eye view: Cooperative exploration by ugv and uav. In *2017 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 247–255. IEEE, 2017.

[21] Kapil Katyal, Katie Popek, Chris Paxton, Phil Burlina, and Gregory D Hager. Uncertainty-aware occupancy map prediction using generative networks for robot navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5453–5459. IEEE, 2019.

[22] Kapil Katyal, Katie Popek, Chris Paxton, Joseph Moore, Kevin Wolfe, Philippe Burlina, and Gregory D Hager. Occupancy map prediction using generative and fully convolutional networks for vehicle navigation. *arXiv preprint arXiv:1803.02007*, 2018.

[23] Kapil D. Katyal, Adam Polevoy, Joseph Moore, Craig Knuth, and Katie M. Popek. High-speed robot navigation using predicted occupancy maps. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5476–5482, 2021.

[24] Kapil D Katyal, Adam Polevoy, Joseph Moore, Craig Knuth, and Katie M Popek. High-speed robot navigation using predicted occupancy maps. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5476–5482. IEEE, 2021.

[25] Jing Yu Koh, Harsh Agrawal, Dhruv Batra, Richard Tucker, Austin Waters, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Simple and effective synthesis of indoor 3d scenes. *arXiv preprint arXiv:2204.02960*, 2022.

[26] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[27] Benjamin Kuipers, Edward A Feigenbaum, Peter E Hart, and Nils J Nilsson. Shakey: from conception to history. *Ai Magazine*, 38(1):88–103, 2017.

[28] Nader Mohamed, Jameela Al-Jaroodi, Imad Jawhar, Ahmed Idries, and Farhan Mohammed. Unmanned aerial vehicles applications in future smart cities. *Technological forecasting and social change*, 153:119293, 2020.

[29] Santhosh K Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy anticipation for efficient exploration and navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 400–418. Springer, 2020.

[30] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3563–3573, 2022.

[31] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14356–14366, 2021.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[33] Manish Saroya, Graeme Best, and Geoffrey A. Hollinger. Online Exploration of Tunnel Networks Leveraging Topological CNN-based World Predictions. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6038–6045, Oct. 2020. ISSN: 2153-0866.

[34] Vishnu Dutt Sharma, Jingxi Chen, Abhinav Shrivastava, and Pratap Tokekar. Occupancy map prediction for improved indoor robot navigation. *arXiv preprint arXiv:2203.04177*, 2022.

[35] Vishnu Dutt Sharma, Jingxi Chen, and Pratap Tokekar. Proxmap: Proximal occupancy map prediction for efficient indoor robot navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, (in press).

[36] Vishnu D Sharma, Maymoonah Toubeh, Lifeng Zhou, and Pratap Tokekar. Risk-aware planning and assignment for ground vehicles using uncertain perception from aerial vehicles. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11763–11769. IEEE, 2020.

[37] Vishnu Dutt Sharma, Lifeng Zhou, and Pratap Tokekar. D2coplan: A differentiable decentralized planner for multi-robot coverage. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3425–3431. IEEE, 2023.

[38] Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. Adversarial

masking for self-supervised learning. In *International Conference on Machine Learning*, pages 20026–20040. PMLR, 2022.

[39] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[40] Minghan Wei, Daewon Lee, Volkan Isler, and Daniel Lee. Occupancy map inpainting for online robot navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8551–8557. IEEE, 2021.

[41] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.