Transferring Visual Knowledge: Semi-Supervised Instance Segmentation for Object Navigation Across Varying Height Viewpoints

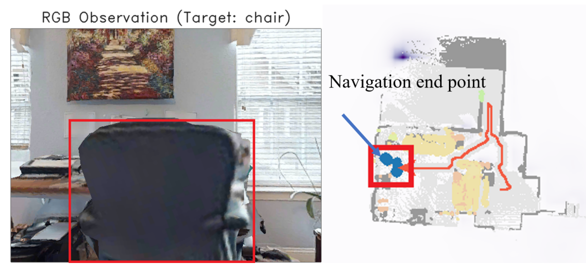
Qiu Zheng, Junjie Hu, *Member, IEEE,* Yuming Liu, Zengfeng Zeng,
Fan Wang and Tin Lun Lam, *Senior Member, IEEE*

***Abstract*— The object navigation task requires robots to understand the semantic regularities in their environments. However, existing modular object navigation frameworks rely on instance segmentation models trained at fixed camera height viewpoints, limiting generalization performance and increasing labeling costs for new height viewpoints. To tackle this issue, we propose a semi-supervised method that transfers knowledge from a source height to a target height, minimizing the need for additional labels. Our approach introduces three key innovations: i) a projection policy to enhance the teacher model's detection capabilities at the target height, ii) a dynamic weight mechanism that emphasizes high-confidence pseudo-labels to reduce overfitting, and iii) a prototype contrast transferring method to transfer knowledge effectively. Experiments on the Habitat-Matterport 3D (HM3D) dataset show our method outperforms state-of-the-art semi-supervised techniques, improving both segmentation accuracy and navigation performance. The code is available at: https://github.com/FreeformRobotics/TransferKnowledge.**

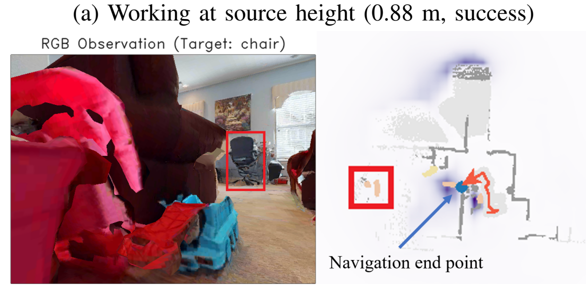
I. INTRODUCTION

Embodied visual navigation enables robots to navigate the physical world using visual inputs [1]. A key task in this domain is object navigation, where an agent must locate a specific object category from a random starting point in an unfamiliar indoor environment [2]. In modular object navigation methods, instance segmentation is crucial for extracting segmentation masks and labels to generate semantic maps.

Previous works [3] [4] show that recognition errors significantly contribute to navigation failures, particularly when visual detection modules are deployed on robots with different height configurations. As illustrated in Fig. 1, instance segmentation models trained at fixed heights struggle with varying viewpoints, leading to reduced navigation success rates when applied to new height configurations. However, supervised learning requires costly and time-consuming data collection and annotation for different camera heights. To address this, we adopt a semi-supervised approach, leveraging existing labeled data and unlabeled data from different heights to enable efficient knowledge transfer.



(a) Working at source height (0.88 m, success)



(b) Working at target height (0.28 m, failure)

Fig. 1: The source model operates at different heights using the Peanut method [5]. The target object (Chair) is highlighted in red boxes, and the navigation stop point is blue. At the target height, the agent mistakenly identifies the sofa as a chair due to the change in viewpoint, resulting to navigation failure.

In this work, we tackle the poor generalization performance of instance segmentation models in object navigation caused by viewpoint differences across varying camera heights. We propose a semi-supervised method that uses labeled data only at the source height and unlabeled data at the target height. Our approach employs a teacher-student training paradigm, where the teacher model generates pseudo-labels for unlabeled target height data, and the student model learns from these pseudo-labels to transfer knowledge across height domains. We introduce three key innovations: i) a projection policy that maps RGB-D images from the source to the target height, enabling the teacher model to gain initial recognition at the target height; ii) a dynamic weight mechanism that prioritizes high-confidence pseudo-labels to reduce noise and overfitting; and iii) a prototype contrast transferring method that uses a prototype memory bank [6] [7] [8]. The memory bank stores and updates prototype centers to facilitate knowledge transfer.

To evaluate our method, we conduct experiments on the

Q.Zheng, J.Hu, Y. Liu and T.L.Lam are with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China.

J.Hu and T.L.Lam are with Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), The Chinese University of Hong Kong, Shenzhen, China.

Z.Zeng and F.Wang are with Baidu Inc., Beijing, China.

Q.Zheng and J.Hu contribute equally.

T.L.Lam is the corresponding author. tllam@cuhk.edu.cn

This work was partially supported by the Guangdong Natural Science Fund under Grant 2024A1515010252, Shenzhen Science and Technology Program under Grants JCYJ20241202124010014 and JCYJ20220818103000001, and the Shenzhen Institute of Artificial Intelligence and Robotics for Society under Grants AC01202101103, and Baidu Inc.

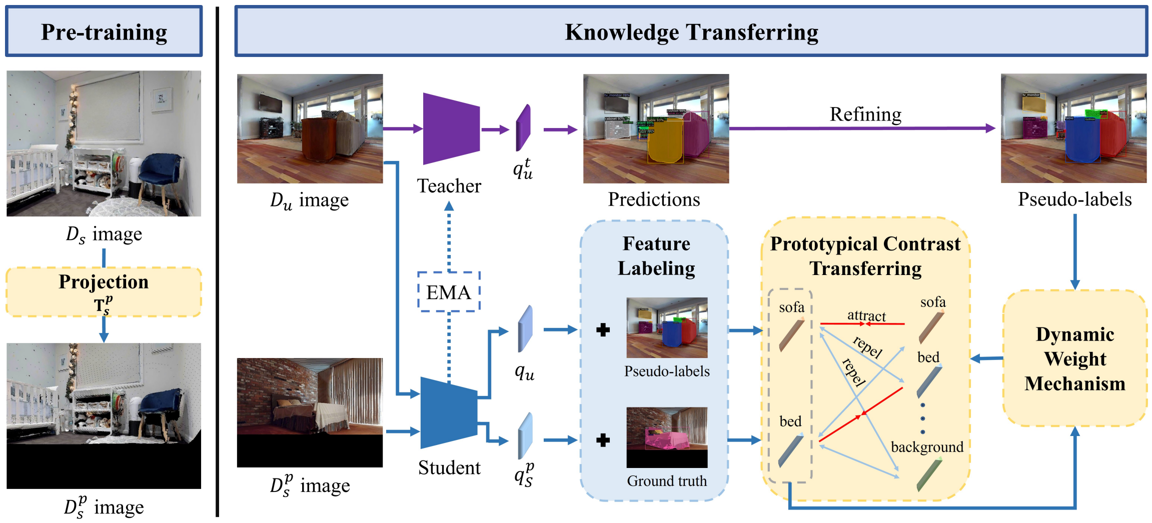


Fig. 2: Framework of our semi-supervised instance segmentation for knowledge transferring. In the pre-training stage, we project RGB-D sensor inputs and ground truth instance labels from source height to target height viewpoint through transformation matrix $\mathbf{T}_s^p$. The projected D_s^p and raw D_s data are used to train the teacher model. In the knowledge transfer stage, the teacher model generates the feature q_u^t of D_u image and set the pseudo-labels. The student model produces the feature q_u and q_s^p of D_u and D_s^p images, respectively. After annotating q_u and q_s^p with the corresponding pseudo-labels and projection ground truth labels, a dynamic weight mechanism allocates weights for each feature during the prototypical contrast transfer process.

HM3D household indoor scenes benchmark [9]. Using an RGB-D dataset with varying camera heights, we demonstrate that our method outperforms state-of-the-art semi-supervised approaches in transferring knowledge to target height viewpoints and significantly improves the navigation success rate of modular object navigation methods at the target height.

Our contributions are summarized as follows:

- We highlight a key limitation of current modular navigation methods: their effectiveness is restricted to homogeneous platforms. Variations in camera heights on heterogeneous robots degrade recognition accuracy and decrease navigation success rate and efficiency.
- We propose a novel semi-supervised method that utilizes a teacher-student paradigm, featuring a projection policy, a dynamic weight mechanism, and a prototype contrast transferring method address the viewpoint knowledge transferring problem.
- Our approach outperforms existing semi-supervised methods in mask-AP and navigation success rates.
- We collected and released an instance segmentation dataset for HM3D household scenes, including RGB-D images with instance masks and semantic maps collected at different camera heights.

II. RELATED WORK

A. Object Goal Navigation

Object navigation is a challenging area in embodied AI, primarily divided into end-to-end and modular approaches [10]. End-to-end methods use a single neural network to encode visual and location information, generating low-level actions without explicit 2D maps. This enables navigation in complex multi-floor environments but relies on large datasets [11] and struggles with generalization. Recently, ResNet [11] [12], ViT [13] [14], and CLIP-based [15] visual encoders have been integrated into navigation networks with some methods employing graph neural networks (GNNs) to improve success rate and search efficiency [16] [17].

In contrast, modular methods segment the system into specialized modules, promoting adaptability to diverse datasets and real-world conditions. This structure enhances generalization during sim-to-real transitions compared to end-to-end methods. However, both explicit [18] [19] [5] and implicit prediction policies [20] [21] [4], as well as those using large language models (LLMs) as prediction policies [22] [23], rely on visual recognition modules that are sensitive to changes in height and viewpoint. Therefore, modular methods often face recognition errors due to varying viewpoints. Our proposed semi-supervised method aims to transfer the performance of the modular approach to robots of different heights in a low-cost manner.

B. Semi-supervised Instance Segmentation

Instance segmentation combines object detection and semantic segmentation, utilizing single-stage or two-stage detectors. Single-stage methods [24] [25] [26] [27] offer faster processing but lower precision due to the absence of a fine processing step. For two-stage methods, Mask R-CNN [28], a typical work, added a mask head into Faster R-CNN [29] to segment objects. Recent advancements include the vision transformer (ViT) [30], which has shown effectiveness in image classification, and [31] [32] demonstrate strong performance in instance segmentation.

Semi-supervised learning techniques leverage both labeled and unlabeled data, commonly using a teacher-student framework for instance segmentation [33] [34] [35] [36] [37]. In this framework, a teacher model annotates unlabeled images, enabling a student model to learn from both labeled and pseudo-labeled data. Guided [33] applies the framework to vision transformer architectures and uses bipartite matching for pairing student proposals with pseudo instances, while PAIS [37] uses mask head scores as loss weights for taking advantage of object masks with different qualities. However, these methods frequently neglect the challenges of height viewpoint changes in unlabeled data, leading to inadequate knowledge transfer. Our work aims to bridge this gap through a cost-effective semi-supervised approach, enhancing robot performance at the target height.

C. Contrastive Learning

Contrastive learning [38] [39] [8] is a self-supervised strategy that maximizes similarity between positive samples while minimizing it for negative ones. This approach is commonly employed to enhance the differentiation between various feature categories [40] [7] [41]. However, the application in instance segmentation models remains limited. Our method integrates the principles of contrastive learning with soft labels to facilitate more effective knowledge transfer.

III. METHOD

Formally, let $D_s = \{x_n^s, d_n^s, y_n^s\}_{n=1}^N$ as a labeled set with N samples collected for training an instance segmentation model at source height where x_n, d_n, and y_n denote an image, depth map, and ground truth instance label, respectively, and $D_u = \{x_m^u\}_{m=1}^M$ as an unlabeled set with M samples at target height. We aim to learn a model that can accurately perform instance segmentation at the target height.

Our semi-supervised instance segmentation framework includes i) a teacher pre-training stage and ii) a teacher-to-student knowledge transfer stage, as shown in Fig. 2. In the first stage, we project data from the source height to the target height to mitigate significant viewpoint differences when training the teacher model. In the second stage, we use the dynamic weight mechanism to assign weights for pseudo-labels, and the weights will mitigate the impact of wrong pseudo-labels on the student model during prototypical contrast transferring progress.

A. Teacher Per-training Stage

To mitigate the significant viewpoint difference, we first propose projecting D_s from the source height to the target height, assuming the camera remains consistent at both heights. Using the known intrinsic parameters $\mathbf{K}$ of the camera and depth images, we map annotated image pixels into the source height camera coordinate system as 3D point $\mathbf{P}^s$. The projection policy is defined as follows:

$$\mathbf{P^p} = \mathbf{T_s^p P^s} \tag{1}$$

where $\mathbf{P}^p$ represents the 3D projection point with its annotation in the target height camera coordinate system, and $\mathbf{T}_s^p$ denotes the relative pose between the two coordinate systems. We then map $\mathbf{P}^p$ to the target height viewpoint image using $\mathbf{K}$, resulting in D_s^p, the data after projection. An example of this projection is shown in Fig. 2.

Then, we train the teacher model using both D_s and D_s^p following the training loss of Mask R-CNN [28].

B. Knowledge Transfer Stage

In this stage, we adopt the teacher-student framework for semi-supervised learning. The student model initializes with parameters from the pre-trained teacher model. Initially, the teacher model remains fixed but is gradually unfrozen during training, with its parameters updated via an exponential moving average (EMA) of the student model. During this stage, we only utilize the datasets D_s^p and D_u. In each iteration, the teacher model generates and refines pseudo-labels for D_u dataset, then the dynamic weight mechanism is applied for prototype memory bank updates and prototypical contrast transferring.

The knowledge transfer process centers on a prototype memory bank, which stores prototype centers for C categories. These prototype centers dynamically shift to facilitate knowledge transfer from the source to the target height viewpoint. Prior to semi-supervised training, we initialize the prototype memory bank using D_s^p dataset to ensure the prototype centers align closer to the target height domain. The class prototype center Q_c is initialized as:

$$Q_c = \frac{\sum_{b=1}^B \mathbf{1}[y_b = c] q_b}{\sum_{b=1}^B \mathbf{1}[y_b = c]} \tag{2}$$

where q_b is the b-th extracted proposal feature vector, c is the category index, and B is the total number of q_b, and $\mathbf{1}[y_b = c]$ is an indicator function checking if the label y_b of q_b matches c.

1) Dynamic Weight Mechanism: During semi-supervised training, many semi-supervised instance segmentation methods use a threshold to filter the teacher model's outputs, which may retain incorrect labels. Therefore, we first refine the pseudo-labels generated by the teacher model for unlabeled data in the following two steps:

- **Filter step**: The teacher model selects candidate instances with maximum class probability exceeding 0.6 and the predicted mask pixel size greater than 200. Meanwhile, the ratio of the mask pixel size to the bounding box size, as well as the aspect ratio of the bounding box, must be within a reasonable range.
- **Merging step**: Overlapping candidate instances are fused based on mask-IOU and box-IOU. The fused instances are defined as the k-th instance pseudo-label $y_{M_k}^u = (y_{M_k^{score}}^u, y_{M_k^{box}}^u, y_{M_k^{mask}}^u)$ in the image, where $y_{M_k^{score}}^u$ contains normalized confidence scores of all fused instances, $y_{M_k^{mask}}^u$ is the binary mask after fusion, and $y_{M_k^{box}}^u$ is the bounding box of $y_{M_k^{mask}}^u$.

To further mitigate the impact of erroneous pseudo-labels, we introduce weight allocation based on confidence scores. The minimum value of $y_{M_k^{score}}^u$ is mapped to a range of $[w_{min}, 1]$ to determine the pseudo-label weight $W_{M_k}^u$, which is also applied in the prototypical contrast transferring method. The value w_{min} increases over training epochs, while ground truth label weights are fixed at 1. During training, these weights are applied to their respective proposal features. The prototype memory bank is updated using the following equation:

$$Q_c^{n_e} = \alpha Q_c^{n_e-1} + (1-\alpha) Q_{c(mini\ batch)}^{n_e} \tag{3}$$

where n_e represents the current iteration, $Q_c^{n_e}$ is the current prototype center for class c and α is a hyperparameter representing the constant update rate. $Q_{c(mini\ batch)}^{n_e}$ denotes the current mini-batch prototype centers, calculated using equation (2), which uses both labeled and unlabeled data features with weights greater than 0.85.

2) Prototypical Contrast Transferring: For D_s^p dataset, we adopt the InfoNCE loss [42] to enhance the distinctiveness of instance features across different classes. The InfoNCE loss $L_{infoNCE}$ is represented as:

$$L_{infoNCE} = -\frac{1}{N_s} \sum_{n_s}^{N_s} \log \left[\frac{e^{\langle \hat{q}_{n_s}, Q_{n_s} \rangle / \tau_1}}{\sum_{c=1}^{C} e^{\langle \hat{q}_{n_s}, Q_c \rangle / \tau_1}} \right] \tag{4}$$

where N_s represents the total number of prediction proposals $\hat{q}_{n_s}$ of D_s^p in the mini-batch. Q_{n_s} is the corresponding prototype center of the proposal feature $\hat{q}_{n_s}$. $\langle \cdot, \cdot \rangle$ denotes cosine similarity between features, with the temperature coefficient set to $\tau_1 = 0.5$.

For D_u dataset, we compute the similarity between features and each prototype center, forming the similarity vector $F_{m_u} = [f_{m_u}^1, \ldots, f_{m_u}^C]$ of the m_u-th prediction proposal feature $\hat{q}_{m_u}$, where C is the number of classes, $f_{m_u}^c$ is defined as follows, with τ_2 the temperature coefficient set to 0.8:

$$f_{m_u}^c = \frac{e^{\langle \hat{q}_{m_u}, Q_c \rangle / \tau_2}}{\sum_{c=1}^{C} e^{\langle \hat{q}_{m_u}, Q_c \rangle / \tau_2}} \tag{5}$$

Furthermore, we construct a similarity matrix S using the prototype memory bank, where each element is defined as $S_{ij} = \langle Q_i, Q_j \rangle$, representing the cosine similarity between the prototype center of classes i and j. Then, we calculate the soft labels for both feature cross-entropy loss and classification cross-entropy loss using S. The feature soft pseudo-label $\tilde{\boldsymbol{F}}_{m_u}$ and class soft pseudo-label $\tilde{\boldsymbol{y}}_{m_u}$ of $\hat{q}_{m_u}$ are defined as:

$$\begin{cases} \tilde{\boldsymbol{F}}_{m_u} = \epsilon_1 \boldsymbol{y}_{m_u} + (1-\epsilon_1) \boldsymbol{y}_{m_u} S \\ \tilde{\boldsymbol{y}}_{m_u} = \epsilon_2 \boldsymbol{y}_{m_u} + (1-\epsilon_2) \boldsymbol{y}_{m_u} S \end{cases} \tag{6}$$

where $\boldsymbol{y}_{m_u}$ is the one-hot label for the class with the highest confidence score in $y_{M_k^{score}}^u$. The values ϵ_1 and ϵ_2 are hyperparameters. The weights $W_{M_k}^u$ assigned to $\hat{q}_{m_u}$ makes the model prioritize the more reliable pseudo-labels. The feature cross-entropy loss L_{contra} and the classification loss L_{ce}^u are given as:

$$L_{contra} = -\frac{1}{M_u} \sum_{m_u}^{M_u} W_{m_u}^u \tilde{\boldsymbol{F}}_{m_u} \log F_{m_u} \tag{7}$$

$$L_{ce}^u = -\frac{1}{M_u} \sum_{m_u}^{M_u} W_{m_u}^u \tilde{\boldsymbol{y}}_{m_u} \log \hat{y}_{m_u} \tag{8}$$

where M_u represents the number of proposals for unlabeled images in the mini-batch, $W_{m_u}^u$ is the corresponding

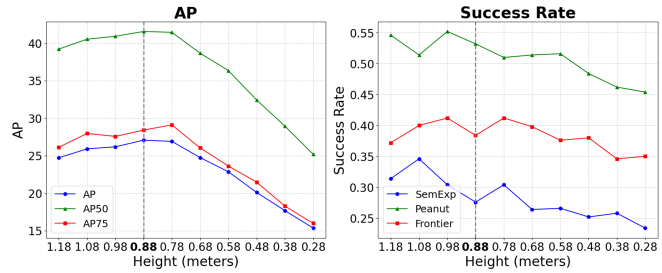


Fig. 3: The changing trend of AP and navigation success rate of source model (0.88 m) at different viewpoints.

weight of m_u-th proposal, and $\hat{y}_{m_u}$ denotes the prediction classification probability.

3) Loss: In the section III-B.2, the supervised loss L_s for dataset D_s^p and the unsupervised loss L_u for dataset D_u are improved based on the training loss of Mask R-CNN [28] and defined as:

$$L_s = L_{ce}^s + \lambda L_{infoNCE} + L_{box} + L_{mask} \tag{9}$$

$$L_u = L_{ce}^u + \lambda L_{contra} + L_{box} + L_{mask} \tag{10}$$

where L_{ce}^s is the classification cross-entropy loss, L_{box} is the bounding-box loss and L_{mask} is the average binary cross-entropy loss, as defined in [28]. The λ is the hyperparameter.

Our optimization objective is to minimize the loss defined as:

$$L = \lambda_s L_s + \lambda_u L_u \tag{11}$$

where λ_s and λ_u are determined by the ratios of labeled and unlabeled batch sizes to the total batch size.

IV. Experiments

In this section, we compare our method with two previous state-of-the-art semi-supervised baselines [37] [33] and evaluate the performance of instance segmentation models trained by these semi-supervised methods on different modular object navigation methods [5] [20] [21].

A. Experimental Setup

1) Dataset: We conduct experiments on a dataset with instance-level segmentation for 21 categories of common household objects, manually collected from the HM3D dataset [9] which is one of the most widely used datasets, and covering as many objects as possible. At each height, we gather 5,548 RGB-D images from 80 training scenes and 991 from 20 validation scenes, spanning heights from 0.28 m to 1.18 m, with increments of 0.1 m. The image size is 480×640. Images captured at 0.88 m serve as labeled source data, while images at target heights are set as unlabeled.

2) Implementation Details: We implement our method using the Detectron2 [43] on a single NVIDIA RTX 3080 GPU with 10 GB. Experiments are tested on Mask R-CNN [28] with R101-FPN backbone. We also evaluate our model on various object navigation modular frameworks [5] [20] [21] using Habitat platform [44]. We adopt a Stochastic Gradient Descent (SGD) optimizer with a base learning rate of 0.0005, a weight decay of 0.05, and early stopping. The learning rate is multiplied by 0.95 per epoch.

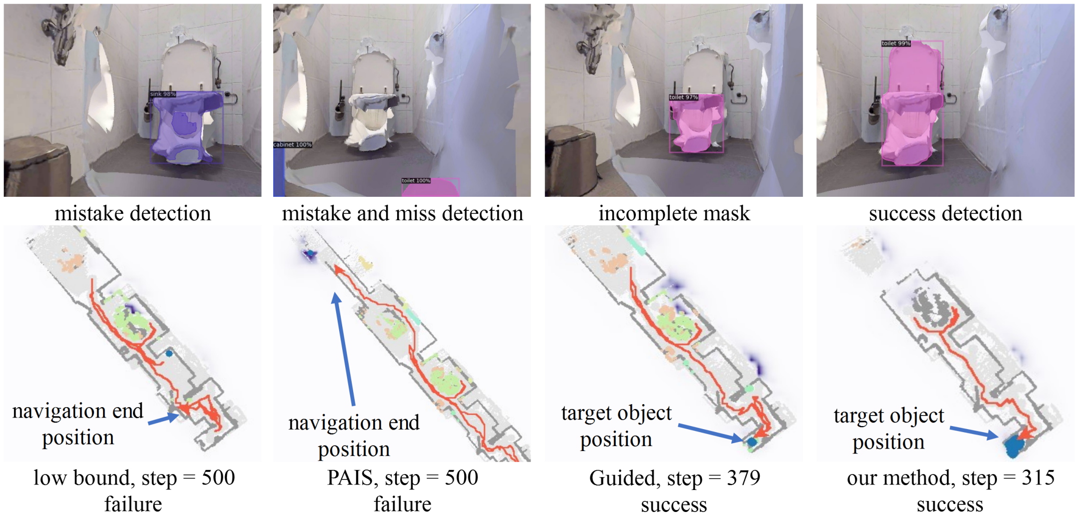


Fig. 4: Visualization results of Peanut navigation method [5] in a scene from HM3D (val). The top row shows the agent's RGB observations with various detection outcomes for the goal object (toilet). The bottom row presents the corresponding semantic maps at the end of episode, indicating navigation end positions (red arrow) and target object positions, along with the number of steps taken for each instance segmentation model training method.

3) Evaluation Metrics: To assess instance segmentation performance, we use the mask-AP metric [28]. For the object navigation task, we follow the evaluation method [2], employing Success Rate (SR) and Success weighted by Path Length (SPL).

B. Analysis of The Influence of Height Viewpoint Variation

We evaluate the source height model (0.88 m) from various viewpoints. As shown in Fig. 3, the average precision (AP) reaches its maximum near the source height and gradually decreases as the height moves away from it. Similarly, the navigation success rates of SemExp [20], Peanut [5], and Frontier [21] exhibit this trend. Notably, at lower heights, changes in viewpoint significantly impact the navigation success rate.

C. Main Results

1) Semi-supervised Instance Segmentation Results: In Table I, we present results from a dataset collected from HM3D scenes, comparing the knowledge transfer performance of two semi-supervised approaches: PAIS [37] and Guided [33]. We also include a lower bound model trained solely on the D_s dataset and an upper bound model trained on both D_s and D'_u (which consists of D_u with ground truth labels).

The results indicate that our method surpasses the other semi-supervised techniques across various transfer heights. For example, at the target height of 0.68 m, our method achieves an impressive improvement in AP50, rising from 38.66 to 42.07 (+3.41), significantly outperforming the PAIS method which scored 40.97 and the Guided method with 40.89.

A closer analysis of performance across different target heights reveals that as the transfer height increases, the improvement in model accuracy becomes more pronounced. Notably, at the 0.28 m target height, our method records 21.10 points AP and 34.15 points AP50, indicating a substantial capability for knowledge transfer.

2) The Results of Transfer Model on Object Navigation Method: Table II presents the results of various modular object navigation methods [20] [5] [21] across different transfer target heights (0.28 m, 0.48 m, and 0.68 m). Additionally, we apply ground-truth object segmentation (GT) in these navigation methods.

The data demonstrate that our approach enhances both the navigation success rate and efficiency across all three object navigation methods and three specified target heights. For instance, within the Peanut [5], our method achieved an increase in the success rate from 0.514 to 0.544 and improved the SPL from 0.261 to 0.265 at the height of 0.68 m, outperforming both the lower bound and other semi-supervised techniques.

Similarly, in the Frontier [21] at a height of 0.68 m, our method maintained a competitive success rate of 0.434 and an SPL of 0.204, demonstrating its effectiveness in improving navigation outcomes through the transfer of knowledge from source to target heights. Notably, as the distance of transferring height increases, the success rate of our method shows significant improvement, with a maximum enhancement of 0.128 for the SemExp [20], rising from 0.234 to 0.362.

These results indicate the potential of our semi-supervised approach in enhancing object navigation performance across varying target heights. The navigation visualization results are shown in Fig. 4.

D. Ablation Study

In this section, we conduct ablation experiments on the model that transfers knowledge from the height of 0.88 m to 0.28 m. To evaluate the significance and functionality of the different sub-methods, we consider the following configurations:

TABLE I: Results of Semi-supervised model, the source height is 0.88 m. D'_u denotes the labeled version of D_u.

Method	Dataset	Transfer Target Height								
		0.28m			0.48m			0.68m		
		AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75
Mask-RCNN [28], lower bound	D_s	15.34	25.18	15.97	20.09	32.42	21.46	24.74	38.66	26.03
Mask-RCNN [28], upper bound	$D_s \bigcup D'_u$	23.57	37.54	24.30	25.10	39.75	26.23	26.26	41.62	27.38
PAIS [37]	$D_s \bigcup D_u$	16.75	28.07	17.68	21.29	35.53	22.19	25.71	40.97	27.11
Guided [33]	$D_s \bigcup D_u$	16.57	27.79	17.04	21.35	35.08	22.93	25.83	40.89	27.52
Our Method	$D_s \bigcup D_u$	**21.10**	**34.15**	**22.02**	**23.53**	**38.34**	**24.57**	**26.66**	**42.07**	**27.99**

TABLE II: Results on object navigation.

Navigation	Method	Target Height					
		0.28m		0.48m		0.68m	
		SR	SPL	SR	SPL	SR	SPL
SemExp [20]	GT	0.368	0.187	0.392	0.197	0.350	0.177
	lower bound	0.234	0.119	0.252	0.123	0.264	0.133
	PAIS [37]	0.228	0.110	0.252	0.125	0.280	0.135
	Guided [33]	0.238	0.111	0.268	0.137	0.264	0.132
	Our Method	**0.362**	**0.162**	**0.300**	**0.145**	**0.328**	**0.140**
Peanut [5]	GT	0.704	0.341	0.676	0.351	0.710	0.374
	lower bound	0.454	0.204	0.484	0.237	0.514	0.261
	PAIS [37]	0.442	0.198	0.496	0.242	0.512	0.264
	Guided [33]	0.448	0.202	0.484	0.236	0.490	0.251
	Our Method	**0.552**	**0.244**	**0.540**	**0.257**	**0.544**	**0.265**
Frontier [21]	GT	0.570	0.294	0.568	0.297	0.566	0.303
	lower bound	0.350	0.180	0.380	0.187	0.398	0.200
	PAIS [37]	0.312	0.155	0.312	0.164	0.370	0.199
	Guided [33]	0.324	0.169	0.336	0.177	0.376	0.198
	Our Method	**0.392**	**0.196**	**0.428**	**0.195**	**0.434**	**0.204**

TABLE III: Results of Ablation Study.

Method				AP	Success Rate		
Projection	Weight Allocation	Prototype	Soft Label		SemExp	Peanut	Frontier
				15.34	0.234	0.454	0.350
✓				19.60	0.272	0.508	0.350
✓	✓			20.45	0.296	0.516	0.390
✓	✓	✓		21.21	0.268	0.510	0.360
✓	✓	✓	✓	21.10	0.362	0.552	0.392

- **Projection only**: The model is trained solely using the labeled images D_s and D_s^p.
- **Projection + Weight Allocation**: This configuration involves pre-training the model using the labeled images D_s and D_s^p, and only using weight allocation in the knowledge transfer stage.
- **Projection + Weight Allocation + Prototype**: Add the prototype memory bank based on the Projection + Weight Allocation item.
- **Projection + Weight Allocation + Prototype + Soft Label**: This represents our complete methodology, as described in Section III.

The significance and functionality of each sub-method are summarized in Table III. Results indicate that the projection method initially transfers knowledge from the source height to the target height and improves object navigation success rates. Incorporating weight allocation further enhances recognition accuracy and overall success. The model needs to utilize the projection method to transfer the knowledge,

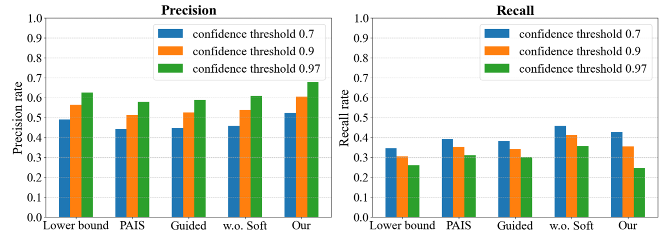


Fig. 5: The analysis of precision and recall rates of three different confidence score threshold (0.7, 0.9, 0.97). We compare the models including Lower bound, PAIS [37], Guided [33], w.o. Soft (Projection + Weight Allocation + Prototype) and our method model.

while weight allocation helps mitigate the effects of incorrect pseudo-labels.

Ablation studies on the prototype memory bank and soft label reveal that using the memory bank alone only slightly improves Average Precision (AP). However, its combination with soft labels is essential for significantly enhancing navigation success rates.

Notably, the problem of navigation success rate tends to decrease as mask-AP increases, a trend also observed in the PAIS [37] and Guided [33] semi-supervised methods, as shown in Table I and II. We analyze the precision and recall rates across different confidence thresholds of 0.7, 0.9, and 0.97. In Fig. 5, the results reveal that PAIS, Guided, and our method without soft labels (Projection + Weight Allocation + Prototype) effectively improve recall rates, however, the presence of incorrect pseudo-labels reduces precision. In contrast, our method adjusts model confidence to an optimal range for object navigation, significantly enhancing the navigation success rate.

V. CONCLUSIONS

In this work, we quantitatively demonstrate that changes in camera height viewpoints impair recognition accuracy and reduce navigation success rates and efficiency. To address this, we propose a semi-supervised method that enables cost-effective knowledge transfer for instance segmentation models from a source height to a target height. Through extensive experiments on HM3D scenes, we show that our method significantly improves the performance of instance segmentation models at the target height and enhances the success rate and efficiency of modular object navigation frameworks. Ablation studies further validate the importance and functionality of each sub-method, highlighting the need for vision modules with an appropriate prediction confidence range in modular object navigation frameworks.

References

[1] M. Deitke, D. Batra, Y. Bisk, T. Campari, A. X. Chang, D. S. Chaplot, C. Chen, C. P. D'Arpino, K. Ehsani, A. Farhadi, L. Fei-Fei, A. Francis, C. Gan, K. Grauman, D. Hall, W. Han, U. Jain, A. Kembhavi, J. Krantz, S. Lee, C. Li, S. Majumder, O. Maksymets, R. Martín-Martín, R. Mottaghi, S. Raychaudhuri, M. Roberts, S. Savarese, M. Savva, M. Shridhar, N. Sünderhauf, A. Szot, B. Talbot, J. B. Tenenbaum, J. Thomason, A. Toshev, J. Truong, L. Weihs, and J. Wu, "Retrospectives on the embodied ai workshop," 2022.

[2] 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," *CoRR*, vol. abs/2006.13171, 2020.

[3] R. Ramrakhya, D. Batra, E. Wijmans, and A. Das, "Pirlnav: Pretraining with imitation and rl finetuning for objectnav," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 17 896–17 906.

[4] J. Wasserman, G. Chowdhary, A. Gupta, and U. Jain, "Exploitation-guided exploration for semantic embodied navigation," 2023.

[5] A. J. Zhai and S. Wang, "Peanut: Predicting and navigating to unseen targets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 10 926–10 935.

[6] H. Wang, Y. Wang, Y. Shen, J. Fan, Y. Wang, and Z. Zhang, "Using unreliable pseudo-labels for label-efficient semantic segmentation," 2023.

[7] Z. Jiang, Y. Li, C. Yang, P. Gao, Y. Wang, Y. Tai, and C. Wang, "Prototypical contrast adaptation for domain adaptive semantic segmentation," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 36–54.

[8] K. Wei, X. Yang, Z. Xu, and C. Deng, "Class-incremental unsupervised domain adaptation via pseudo-label distillation," *IEEE Transactions on Image Processing*, vol. 33, pp. 1188–1198, 2024.

[9] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. M. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and D. 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*, 2021.

[10] T. Gervet, S. Chintala, D. Batra, J. Malik, and D. S. Chaplot, "Navigating to objects in the real world," *Science Robotics*, vol. 8, no. 79, p. eadf6991, 2023.

[11] R. Ramrakhya, E. Undersander, D. Batra, and A. Das, "Habitat-web: Learning embodied object-search strategies from human demonstrations at scale," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5163–5173.

[12] K. Yadav, R. Ramrakhya, A. Majumdar, V.-P. Berges, S. Kuhar, D. Batra, A. Baevski, and O. Maksymets, "Offline visual representation learning for embodied navigation," in *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023.

[13] K. Yadav, A. Majumdar, R. Ramrakhya, N. Yokoyama, A. Baevski, Z. Kira, O. Maksymets, and D. Batra, "Ovrl-v2: A simple state-of-art baseline for imagenav and objectnav," 2023.

[14] Y. Hong, Y. Zhou, R. Zhang, F. Dernoncourt, T. Bui, S. Gould, and H. Tan, "Learning navigational visual representations with semantic map supervision," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3032–3044.

[15] H. Yoo, Y. Choi, J. Park, and S. Oh, "Commonsense-aware object value graph for object goal navigation," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4423–4430, 2024.

[16] Y. Lyu, Y. Shi, and X. Zhang, "Improving target-driven visual navigation with attention on 3d spatial relationships," *Neural Processing Letters*, vol. 54, no. 5, pp. 3979–3998, Oct 2022.

[17] M. Lingelbach, C. Li, M. Hwang, A. Kurenkov, A. Lou, R. Martín-Martín, R. Zhang, L. Fei-Fei, and J. Wu, "Task-driven graph attention for hierarchical relational object navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 886–893.

[18] J. Wang and H. Soh, "Probable object location (polo) score estimation for efficient object goal navigation," 2023.

[19] M. Zhu, B. Zhao, and T. Kong, "Navigating to objects in unseen environments by distance prediction," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 10 571–10 578.

[20] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 4247–4258.

[21] B. Yu, H. Kasaei, and M. Cao, "Frontier semantic exploration for visual target navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 4099–4105.

[22] K. Zhou, K. Zheng, C. Pryor, Y. Shen, H. Jin, L. Getoor, and X. E. Wang, "ESC: Exploration with soft commonsense constraints for zero-shot object navigation," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 42 829–42 842.

[23] W. Cai, S. Huang, G. Cheng, Y. Long, P. Gao, C. Sun, and H. Dong, "Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill," 2023.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.

[27] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[28] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[31] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 17 864–17 875.

[32] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 1290–1299.

[33] T. Berrada, C. Couprie, K. Alahari, and J. Verbeek, "Guided distillation for semi-supervised instance segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 475–483.

[34] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, "Unbiased teacher for semi-supervised object detection," 2021.

[35] Y.-C. Liu, C.-Y. Ma, and Z. Kira, "Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 9819–9828.

[36] D. Filipiak, A. Zapała, P. Tempczyk, A. Fensel, and M. Cygan, "Polite teacher: Semi-supervised instance segmentation with mutual learning and pseudo-label thresholding," *IEEE Access*, vol. 12, pp. 37 744–37 756, 2024.

[37] J. Hu, C. Chen, L. Cao, S. Zhang, A. Shu, G. Jiang, and R. Ji, "Pseudo-label alignment for semi-supervised instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 16 337–16 347.

[38] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings*

of the 37th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607.

[39] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[40] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4248–4257.

[41] Q. Wei, L. Feng, H. Sun, R. Wang, C. Guo, and Y. Yin, "Fine-grained classification with noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 11 651–11 660.

[42] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019.

[43] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[44] Manolis Savva*, Abhishek Kadian*, Oleksandr 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.