2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

A Contrastive Few-shot RGB-D Traversability Segmentation Framework for Indoor Robotic Navigation

Qiyuan An[1,2], Tuan Dang[3], and Fillia Makedon[1]

***Abstract*— Indoor traversability segmentation aims to identify safe, navigable free space for autonomous agents, which is critical for robotic navigation. Pure vision-based models often fail to detect thin obstacles, such as chair legs, which can pose serious safety risks. We propose a multi-modal segmentation framework that leverages RGB images and sparse 1D laser depth information to capture geometric interactions and improve the detection of challenging obstacles. To reduce the reliance on large labeled datasets, we adopt the few-shot segmentation (FSS) paradigm, enabling the model to generalize from limited annotated examples. Traditional FSS methods focus solely on positive prototypes, often leading to overfitting to the support set and poor generalization. To address this, we introduce a negative contrastive learning (NCL) branch that leverages negative prototypes (obstacles) to refine free-space predictions. Additionally, we design a two-stage attention depth module to align 1D depth vectors with RGB images both horizontally and vertically. Extensive experiments on our custom-collected indoor RGB-D traversability dataset demonstrate that our method outperforms state-of-the-art FSS and RGB-D segmentation baselines, achieving up to 9% higher mIoU under both 1-shot and 5-shot settings. These results highlight the effectiveness of leveraging negative prototypes and sparse depth for robust and efficient traversability segmentation.**

I. Introduction

Traversability segmentation is a fundamental task in robotic navigation, aiming to identify freespace that is safe for autonomous agents to traverse. While most existing work focuses on outdoor scenarios such as self-driving cars [1], [2], indoor environments remain relatively underexplored despite their practical importance in warehouse automation, hotel service, and hospital robotics [3], [4]. Compared to outdoor scenes that often feature structured roads and lane markings, indoor traversability segmentation poses unique challenges due to varying lighting conditions, complex floor textures, cluttered layouts, and the presence of arbitrarily moving humans or objects [5].

Most prior work adopts purely vision-based solutions [6], [7], [8]. However, our analysis shows that even state-of-the-art segmentation models, such as Deeplabv3+ [9] and SegFormer [10], struggle to detect thin obstacles such as chair legs. Although these objects occupy only a small fraction of image pixels and have minimal influence on global segmentation metrics, failing to detect them can pose significant safety risks to robots.

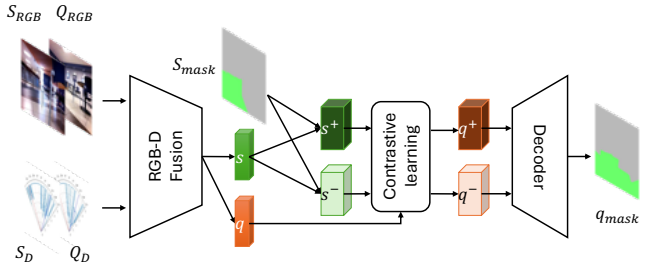

Fig. 1: Illustration of our RGB-D few-shot segmentation framework. The support and query inputs consist of RGB images (S_{RGB}, Q_{RGB}) and depth vectors (S_D, Q_D), encoded separately and fused via a multi-modal fusion block to produce support and query features (s, q). The support mask S_{mask} is used to mask-pool on s, yielding both positive and negative prototypes (s^+, s^-). The query feature q is then refined into free-space and obstacle representations (q^+, q^-), which are concatenated and digested by a lightweight decoder to generate the final query mask (q_{mask}).

Unlike conventional RGB-D datasets such as NYUv2 [11] or SUN RGB-D [12], which provide dense 2D depth maps registered with corresponding RGB images, our custom-collected dataset consists of paired RGB images and 1D laser scans. We adopt this sensing setup for several reasons. First, many commercial indoor robots (e.g., cleaning, delivery, and assistive robots) are equipped with lightweight, low-cost 1D LiDARs rather than expensive 2D or 3D depth cameras, making our dataset more representative of real-world deployment scenarios. Second, 1D laser sensors enable large-scale, long-duration data collection at a fraction of the cost of dense depth sensors, thereby allowing us to build a large-scale dataset of thousands of RGB–1D depth pairs. Finally, the sparse and partial nature of 1D scans introduces a unique technical challenge: the depth backbone must learn to effectively encode 1D inputs and align them with 2D RGB features, which differs substantially from prior multimodal segmentation frameworks such as DFormer [13] and CANet [14].

Therefore, our dataset is inherently more challenging than conventional RGB-D datasets: the 1D depth signals are vertically degenerated and often unregistered with respect to the RGB images. This setting reflects real-world conditions in which robot-mounted sensors are limited in resolution,

[1]Qiyuan An was with Department of Computer Science and Engineering, University of Texas at Arlington, 1225 West Mitchell, Arlington, TX 76019, USA qxa5560@mavs.uta.edu. He is now with Uber, Sunnyvale, CA 94086 USA qiyuan.an@uber.com.

[3]Tuan Dang is with Cognitive Robotics Lab, Department of Electrical Engineering and Computer Science, University of Arkansas, Fayetteville, AR, USA tuand@uark.edu

[1]Fillia Makedon is with Department of Computer Science and Engineering, University of Texas at Arlington, 1225 West Mitchell, Arlington, TX 76019, USA. makedon@uta.edu

field of view, or calibration accuracy. Consequently, models trained on our dataset must be robust to sensor imperfections, unregistration, and sparse depth information—properties that are critical for practical indoor robotic navigation.

Another challenge comes from the unregistered depth vectors, which do not align well with the vertical beams of the corresponding RGB images in our custom-collected dataset. To address this, we introduce a two-stage attention depth module that dynamically maps the 1D depth to its paired RGB image along both vertical and horizontal dimensions. This design not only eliminates the need for explicit registration but also enables the model to capture dynamic geometric interactions between RGB and depth features.

Training a reliable traversability segmentation model also faces several practical challenges. Chief of them is acquiring large-scale, fine-grained annotations, which is often expensive, time-consuming, and labor-intensive [15]. Few-shot segmentation (FSS) addresses this challenge by enabling models to learn from a limited number of labeled examples (the *support* set) and generalize to new, unseen instances (the *query* set) [16], [17], [18]. We adopt the meta-learning paradigm in which the query set is matched to support prototypes at the feature level, allowing the model to generalize with minimal supervision.

In the context of indoor traversability segmentation, traditional prototype-matching methods first learn a *positive prototype* from the support set, representing traversable freespace. Query pixels that closely match this prototype in the feature space are then classified as freespace. However, this strategy is prone to misclassification when regions share similar textures or colors—for example, confusing white walls with white ceramic floor tiles. Such reliance on positive prototypes alone leads to overfitting to the support set and poor generalization to unseen scenarios.

To address this limitation, we propose to explicitly leverage *negative prototypes*—representations of obstacles that are typically ignored in conventional FSS methods. Our negative contrastive learning (NCL) branch first extracts obstacle prototypes from the support set and then identifies corresponding obstacle regions in the query set. These negative regions are used to refine the free-space masks by explicitly expelling obstacles from potential traversable areas. Finally, a lightweight decoder fuses the outputs of both the positive and negative branches to produce the final segmentation mask.

To summarize, we propose a multi-modal traversability segmentation framework with a novel few-shot training paradigm. Our main contributions are:

1) Multi-modal RGB-D segmentation: Integrating RGB images and 1D depth vectors captures geometric interactions and improves detection of thin obstacles.
2) Two-stage attention depth module: Dynamically aligning depth vectors with RGB images along horizontal and vertical dimensions addresses unregistered depth.
3) Negative contrastive learning: Leveraging neglected negative prototypes enhances generalization and reduces overfitting in free-space prediction.
4) Dataset contribution: We collect and release a large-scale indoor RGB-D traversability dataset with sparse 1D depth annotations, providing a new benchmark for future research in indoor navigation.

Our implementation will be publicly available at `https://github.com/qiyuan53/NCL`.

II. Related Work

Traversability segmentation is a critical component of robotic navigation, as it enables autonomous agents to identify freespace while avoiding obstacles. Most existing approaches rely solely on visual information to extract features and recognize traversable regions [6], [19], [1], [20], [3]. However, vision data alone often proves insufficient in complex environments with irregularly shaped objects such as appliances and furniture. For instance, even state-of-the-art segmentation models such as SegFormer [10] struggles to exclude thin structures (e.g., chair legs) from freespaces, even under fully supervised training.

To address these limitations, researchers have increasingly incorporated depth sensing modalities, including LiDAR [21], [7], [22] and depth cameras [13], [23], to complement RGB information in traversability segmentation. While RGB–D fusion has demonstrated clear benefits, training reliable multi-modal segmentation models typically requires large quantities of fine-grained labeled data, which is costly and time-consuming to obtain.

Few-shot learning (FSL) provides a promising alternative by enabling models to quickly adapt to new scenarios using only a limited number of labeled examples [24], [25]. Extending this paradigm, few-shot segmentation (FSS) adapts models on a small support set containing target classes and then infers masks on the query set. However, conventional FSS methods typically focus only on positive prototype matching—aligning query features with prototypes of the target class—while neglecting the background class [26], [27], [17], [18]. In the context of traversability segmentation, this limitation is particularly problematic: positive prototype matching tends to bias the model toward a specific freespace type seen in the support set (e.g., carpet), while failing to generalize across diverse freespace appearances (e.g., ceramic tiles).

To address this drawback, we introduce a novel negative contrastive learning branch that explicitly expels the background class (obstacles) to refine freespace predictions, improving generalization and robustness in unseen environments. To the best of our knowledge, this is the first work to explore few-shot RGB–1D depth traversability segmentation, bridging multi-modal fusion and FSS in a challenging real-world setting.

III. METHODOLOGY

A. Problem Formulation

We tackle the RGB-D traversability segmentation problem within the few-shot segmentation (FSS) framework, aiming to meta-learn a multi-modal model that can quickly adapt a pretrained segmentation model to unseen indoor scenarios given limited labeled examples. Following recent advances

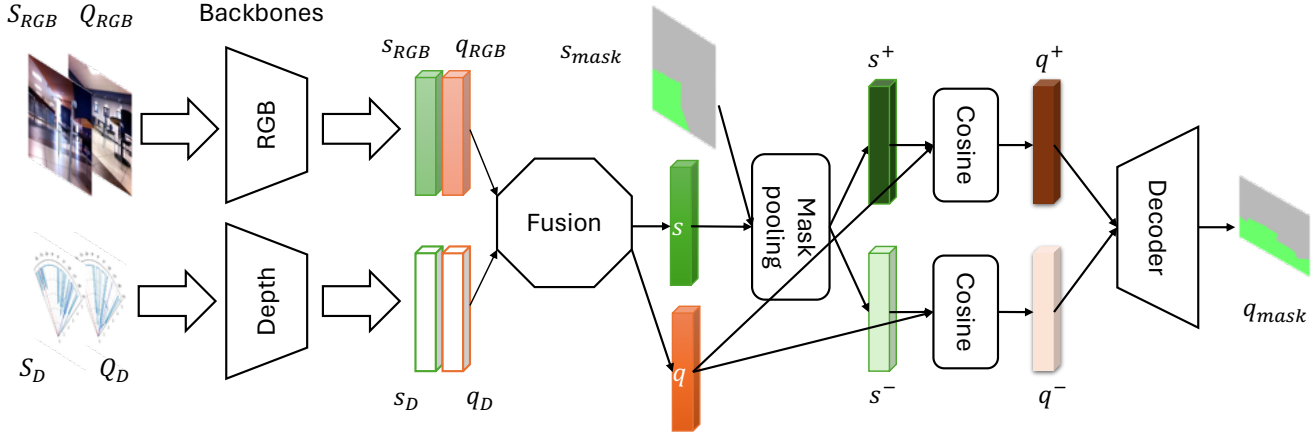

Fig. 2: Proposed contrastive few-shot RGB-D segmentation framework. RGB and depth inputs are embedded with modality-specific backbones, fused into unified support and query features, and refined through prototype-based contrastive learning. A lightweight decoder then predicts the query segmentation mask for indoor freespaces.

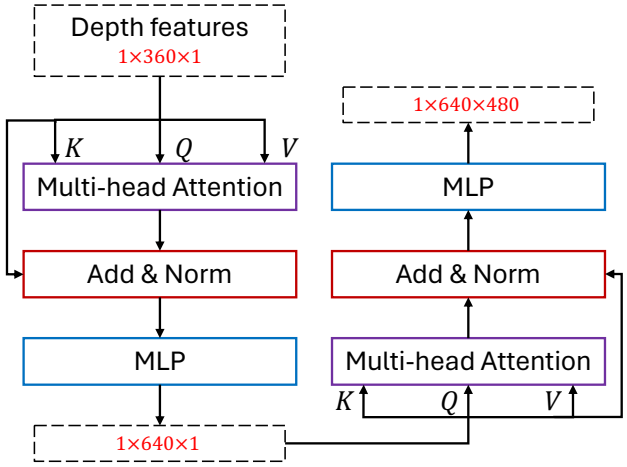

Fig. 3: Two-stage attention depth backbone. It transforms 1D depth vectors into spatially aligned embeddings by applying horizontal attention (beam alignment) followed by vertical attention (height projection), producing refined depth features for multi-modal fusion.

in FSS [27], [17], [28], we adopt an episodic learning protocol for training and evaluation. The training set $\mathcal{D}_{train}$ comprises manually labeled RGB-D pairs, while the test set $\mathcal{D}_{test}$ is drawn from a large collection of unlabeled RGB-D pairs. To generate label masks for $\mathcal{D}_{test}$, we first train a fully supervised segmentation model on $\mathcal{D}_{train}$, achieving a mean Intersection over Union (mIoU) of 98.5% under the 4-fold cross-validation. This high performance indicates its reliability for producing accurate masks used in evaluation. The test set $\mathcal{D}_{test}$ is then inferred on this fully supervised model to produce label masks for later evaluation.

In each training episode, we sample K-shot examples from $\mathcal{D}_{train}$ to form the support set $S = \{S_{RGB}, S_D, S_{mask}\}$, where S_{RGB} and S_D denote the RGB image and depth map, respectively, and S_{mask} is the corresponding ground-truth mask. Similarly, the query set $Q = \{Q_{RGB}, Q_D, Q_{mask}\}$ is drawn from $\mathcal{D}_{test}$. The model adapts to the support set to learn traversable indoor freespaces and infers segmentation masks for the query set. Given the large size of the unlabeled dataset (91,951 RGB-D pairs), 10,000 RGB-D pairs are randomly sampled for evaluation to ensure computational efficiency without compromising statistical validity.

B. Methodology Overview

Our proposed model is designed to efficiently extract and fuse heterogeneous information from RGB and depth modalities while preserving a lightweight training strategy. As illustrated in Figure 2, the model first embeds RGB and depth inputs into latent representations, denoted as $\{s_{RGB}, q_{RGB}\}$ and $\{s_D, q_D\}$, using modality-specific backbones (see Section III-C). These modality-specific features are then integrated through a multi-modal fusion module, yielding unified support and query features, s and q, respectively. The fusion module is flexible in design and can incorporate state-of-art approaches such as [29], [13].

Next, the support feature s is mask-pooled by the support mask s_{mask} to generate positive (s^+) and negative (s^-) prototypes, corresponding to traversable freespace and obstacles. The query feature q is then refined via cosine similarity with s^+ and s^-, producing positive (q^+) and negative (q^-) query features (see Section III-D). Finally, these query features are concatenated and passed to a lightweight decoder, which outputs the final query segmentation mask q_{mask}.

C. RGB and Depth Backbones

The input tuple, consisting of an RGB image and a 1D depth vector, is first embedded by the RGB and depth backbones in parallel. Since our custom-collected RGB images have a resolution of 640×480, we design a lightweight RGB backbone composed of two convolutional layers with 3×3 kernels and a stride of 2, with GeLU activations [30].

To accommodate the format of our depth data—a single-row vector of dimension [360] encoding both distance and viewing angle, we design a novel two-stage attention module

for depth embedding, as illustrated in Figure 3. The robot's laser scanner captures measurements across a horizontal plane, where each pixel corresponds to a beam in the paired image and encodes the distance to obstacles. To avoid explicitly registering each depth value with its corresponding image pixel, the first stage of our depth module applies horizontal self-attention to learn embeddings h_{d1} that align with the beams of the RGB image, as formulated in Eq. 1.

$$\begin{aligned} q_d &= FC(x_d), \quad k_d = FC(x_d), \quad v_d = FC(x_d), \\ h_{d1} &= \text{softmax}\left(\frac{q_d \cdot k_d^T}{\sqrt{m}}\right) \cdot v_d, \end{aligned} \tag{1}$$

where the input depth vector x_d is linearly projected into *Query*, *Key*, and *Value* representations in a self-attention block. The attended feature h_{d1} captures horizontal depth information, where m is the dimension of x_d. This stage also performs depth-to-image registration implicitly.

The second stage vertically attends to the output of the first stage to produce a depth map h_{d2} that matches the image height ($[480]$), as shown in Eq. 2.

$$\begin{aligned} h'_{d1} &= reshape(h_{d1}), \\ q_{d2} &= FC(h'_{d1}), k_{d2} = FC(h'_{d1}), v_{d2} = FC(h'_{d1}), \\ h_{d2} &= \text{softmax}\left(\frac{q_{d2} \cdot k_{d2}^T}{\sqrt{m_2}}\right) \cdot v_{d2}, \end{aligned} \tag{2}$$

where m_2 denotes the dimension of the output from the previous stage (h_{d1}). The resulting features are refined through a standard attention block, consisting of a residual connection with layer normalization and a feed-forward MLP. In this way, our proposed depth module effectively captures the geometric feature both horizontally and vertically and exploits the structural information contained in the 1D depth vector.

Finally, the extracted RGB and depth features are fused using existing multi-modality fusion frameworks such as [29], [13], as defined in Eq. 3.

$$\begin{aligned} s &= Fusion(s_{RGB}, s_D), \\ q &= Fusion(q_{RGB}, q_D). \end{aligned} \tag{3}$$

D. Contrastive Few-shot Learning

Most existing FSS approaches predict the query mask using only positive prototype matching [25], [27], [17], often overlooking the informative role of negative prototypes. A recent method [28] incorporates negative prototypes, but in a parametric manner that introduces additional trainable weights and increases model complexity. To address this limitation, we propose a novel contrastive few-shot learning strategy that fully exploits both positive and negative support prototypes in a non-parametric framework.

Our method comprises two complementary branches. The first branch, positive-to-prototype ($p2p$), follows the traditional prototype matching paradigm [27]: the query's positive feature q^+ is obtained by computing the cosine similarity between the query feature q and the positive support prototype s^+. Query pixels with high similarity to s^+ are assigned to the traversable foreground class.

The second branch, negative-to-prototype ($n2p$), computes the cosine similarity between the query feature and the negative support prototype s^-, producing a negative feature representation q^-. The positive (q^+) and negative (q^-) features are then concatenated and passed to the decoder to predict the final segmentation mask. This contrastive design explicitly models both traversable and non-traversable cues, while introducing no additional parametric overhead.

1) Positive Prototype Matching Branch: The positive prototype matching branch ($p2p$) learns traversable freespaces by following the prototypical network paradigm in few-shot learning [26]. Specifically, it identifies query pixels in q that are most similar to the support set's positive prototype s^+, yielding the positive query feature q^+.

The support mask s_{mask} mask-pools the support feature s, producing positive prototypes s^+ (freespace) and negative prototypes s^- (obstacles), where the mask polarities $+$ and $-$ denote foreground and background, respectively. Mask-pooling, adapted from RoIAlign [31], outperforms traditional global average pooling (GAP) by preserving richer spatial information from the prototype feature map [4].

Finally, the positive query feature q^+ is derived by computing pixel-level cosine similarity between q and s^+, as formulated in Eq. 4:

$$\begin{aligned} s^+ &= mask\_pool(s^+_{mask}, s), \\ q^+ &= cosine(s^+, q). \end{aligned} \tag{4}$$

2) Negative Contrastive Learning Branch: To address the limited generalization ability of conventional positive prototype matching, we propose a Negative Contrastive Learning (NCL) branch ($n2p$). Relying solely on the $p2p$ branch often leads to overfitting to the support set, making it difficult to generalize across diverse freespace appearances (e.g., adapting from dark carpet to white ceramic tiles or colorful plastic floors). This issue is further exacerbated in few-shot settings due to the scarcity of training samples [18], [4].

The $n2p$ branch mitigates this problem by leveraging negative prototypes in a non-parametric manner. Similar to the $p2p$ branch, negative prototypes s^- are derived by s_{mask} mask-pooling on the support feature s. We then compute the cosine similarity between s^- and the query feature q, yielding the negative query feature q^-, which highlights pixels in q resembling obstacles, as shown in Eq. 5:

$$\begin{aligned} s^- &= mask\_pool(s^-_{mask}, s), \\ q^- &= cosine(s^-, q). \end{aligned} \tag{5}$$

We adopt a non-parametric design to keep the $n2p$ branch as generic as possible. Introducing additional learnable layers risks overfitting, since these layers are randomly initialized and trained only on the limited support set in the FSS

protocol. Our experiments confirm this issue, consistent with findings in [28].

Finally, the decoder concatenates both positive (q^+) and negative (q^-) query features to produce the final segmentation mask q_{mask}, as formulated in Eq. 6:

$$q_{mask} = decoder(q^+, q^-). \tag{6}$$

Note that we omit explicit parametric fusing q^+ and q^- like [4], since the decoder already incorporates MLP layers (e.g., as in modern segmentation models such as [13]), making additional ones unnecessary.

E. Other Training Details

We highlight additional training details, given that our RGB-D pipeline incorporates both novel network modules and a new training strategy. To keep the meta-learning process lightweight and minimize learnable parameters, we only update the two-stage depth module and the decoder: the depth module is newly introduced, and the decoder is essential for any segmentation tasks. All other components, including the RGB backbone and fusion blocks, remain frozen. Notably, the mask-pooling and cosine similarity operations used in our training strategy do not introduce any additional learnable parameters.

In summary, our methodology combines lightweight RGB-D backbones, non-parametric contrastive few-shot learning, and flexible multi-modal fusion to achieve robust traversability segmentation under few-shot conditions, while minimizing additional learnable parameters.

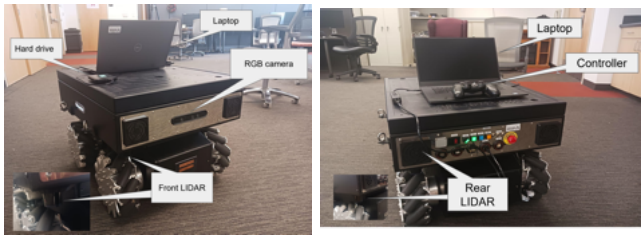

Fig. 4: Summit-XL Steel platform

IV. EXPERIMENTS

A. Dataset Collection

We constructed an indoor traversability segmentation dataset using a tele-operated Summit-XL Steel robotic platform[1], as shown in Figure 4. Multi-modal datastreams were recorded through ROS Melodic[2] with the *message_filters* library[3] to ensure synchronization between RGB images and depth laser data.

The dataset covers multiple campus buildings, including classrooms, cafeterias, corridors, offices, and laboratories, providing diverse indoor layouts, lighting environments and floor materials. In total, the dataset contains 91,951 paired RGB and depth samples, of which 2,553 are manually annotated with freespace masks. Pseudo-labels for the unlabeled portion were generated by a strong fully-supervised teacher model achieving 98.5% mIoU in 4-fold cross-validation, providing high-confidence reference masks. Similar practices can be found in [32].

Table I summarizes the dataset statistics across different buildings / domains. This multi-domain composition ensures a variety of visual and geometric conditions, making the dataset well-suited for evaluating generalization in RGB-D few-shot segmentation.

Building	ELB	ERB	NH	UC
RGB-D pairs	8732	5590	4610	36072
Manually labeled	649	658	433	521
Building	**WH**	**Mocap**	**Heracleia**	**Total**
RGB-D pairs	10899	10872	17295	91951
Manually labeled	292	-	-	2553

TABLE I: Statistics of the custom-collected indoor RGB-D traversability dataset.

B. Metrics and Implementation Details

We report performance using mean Intersection-over-Union (mIoU), the standard metric for semantic segmentation, and adopt an episodic training protocol consistent with prior work [17]. Regarding trainable parameters, the RGB backbone and the multi-modality fusion module from DFormer are frozen [13], while updating only the learnable modules: the proposed depth backbone and the final decoder. The RGB backbone is pretrained on ImageNet-1K [33], and the fusion module is pretrained on NYUDepthv2 [11].

Our dataset is organized into episodes, each consisting of a support set and a query set, where every sample includes an RGB image paired with its corresponding 1D depth vector. We evaluate under both 1-shot and 5-shot configurations, following standard FSS protocols [17], [18], [16]. For each episode, the trainable modules are adapted on the support set for 120 epochs using cross-entropy loss and the AdamW optimizer [34] with an initial learning rate of 6×10^{-5} and scheduled by WarmUpPolyLR with a power of 0.9, a weight decay of 0.01, and 5 warm-up epochs.

C. Quantitative Results

Table II presents the traversability segmentation results on our indoor dataset. To enable comparing with RGB-D segmentation baselines that require 2D depth maps directly, we convert our 1D depth vector into a pseudo-2D map by warping to image width and duplicating along the image's height.

Our method, NCL, consistently outperforms state-of-the-art few-shot segmentation methods across both 1-shot and 5-shot settings. On average, NCL improves mIoU by 5–15 points compared to PANet, CWT, and BAM, with the most significant gains observed in 1-shot scenarios (8 points more than BAM). These improvements are consistent across two

[1] https://robots.ros.org/summit-xl-steel/
[2] https://wiki.ros.org/melodic
[3] http://wiki.ros.org/messagefilters

Backbone	Method	1-shot			5-shot			Trainable / Total Params
		Freespace	Obstacles	mIoU	Freespace	Obstacles	mIoU	
CMX [29]	PANet [27]	74.8	52.01	63.4	76.24	54.81	65.52	2.5M / 65.7M
	CWT [17]	83.51	64.82	74.16	84.87	64.38	74.68	1.2M / 98M
	BAM [28]	86.03	73.39	78.91	85.67	74.69	80.19	2.8M / 102M
	NCL (ours)	91.5	82.55	87.03	91.84	83.94	87.91	4.6M / 59.7M
DFormer [13]	PANet	76.45	53.16	64.8	78.54	56.46	67.5	4.3M / 14.7M
	CWT	85.35	66.25	75.8	87.43	66.8	77.11	1.2M / 46.7M
	BAM	87.93	75.01	81.47	88.26	76.94	82.61	4.9M / 51.6M
	NCL (ours)	93.52	84.37	88.95	94.61	86.5	90.56	4.4M / 29.6M

TABLE II: Quantitative results on the indoor traversability dataset under 1-shot and 5-shot settings. Our proposed NCL achieves the best mIoU across all settings while requiring only a small fraction of trainable parameters, since most backbone weights are frozen during adaptation.

fusion backbones, CMX and DFormer, demonstrating the robustness and generality of our approach. All reported metrics are means over 5 independent runs with different random seeds for episodic sampling; standard deviations were consistently $< 0.4\%$ in mIoU and are omitted for brevity given the large test set size.

The performance gains can be attributed to key design choices. PANet, a traditional prototype-matching method, overlooks negative prototypes and employs global average pooling, thereby losing discriminative information from the support set. CWT transfers only the classifier layer of a transformer-based backbone, limiting its ability to model complex cross-modal interactions. BAM attempts to separate background objects but relies on additional parametric layers, which are prone to overfit under few-shot conditions. In contrast, NCL leverages non-parametric negative contrastive cues, capturing both positive and negative information without extra trainable parameters, resulting in more robust generalization to unseen query scenarios.

These results indicate that explicitly modeling negative prototypes via contrastive learning significantly improves few-shot RGB-D traversability segmentation, making our approach both accurate and backbone-agnostic.

D. Qualitative Results

A primary objective of this work is to overcome the limitations of pure vision-based models in recognizing thin obstacles, such as chair legs. Although these objects often occupy less than 1% of image pixels and have little impact on mIoU, they pose serious safety risks to autonomous agents and surrounding people. To highlight the effectiveness of our method, we present representative qualitative examples in Figure 5.

Each row shows (1) the query RGB image, (2) the corresponding depth vector, where the left and right boundaries of RGB sights are shown in red and blue dashed lines, and values represents the distance to obstacles up to 5 meters, (3) predictions from a model without the two-stage depth attention module and without the NCL branch, (4) predictions with the depth module but without NCL, and (5) predictions from our complete model with both components enabled.

We observe three trends:

1) Without the depth module (col. 3): the model frequently confuses floors with visually similar walls or ceilings, leading to significant false positives.
2) With the depth module only (col. 4): the model better separates floors from walls/ceilings, but still fails to exclude thin objects such as chair legs.
3) Full model with NCL (col. 5): the segmentation improves significantly, successfully excluding thin obstacles and yielding clean, safe freespace predictions.

These results visually confirm the complementary benefits of the proposed two-stage depth attention module and the negative contrastive learning branch.

Backbone	Setting	1-shot		
		Freespace	Obstacles	mIoU
DFormer [13]	$-H$ $-W$	85.84	69.1	77.47
	$-H$ $+W$	88.63	71.44	80.03
	$+H$ $-W$	90.32	76.74	83.53
	$+H$ $+W$	93.52	84.37	88.95

TABLE III: Ablation study on the two-stage depth attention module, where H and W represent horizontal and vertical attention blocks respectively.

E. Ablation Study

1) Two-stage Attention Depth Module: We conduct an ablation study on the proposed two-stage attention depth module using the DFormer backbone under the 1-shot FSS setting. Results are summarized in Table III.

The baseline setting ($-H$ $-W$) directly warps the depth vector to the image width and duplicates it across rows without any attention mechanism. This achieves the lowest performance (77.47 mIoU), as it fails to exploit the geometric cues inherent in the depth signal. Adding only width attention ($-H$ $+W$) improves mIoU by +2.6, since it better aligns depth values with horizontal pixel positions. Introducing only height attention ($+H$ $-W$) achieves a larger gain of +6.1 over the baseline, highlighting the importance of vertical depth cues. For instance, a sudden decrease in depth often corresponds to nearby obstacles, directly constraining the freespace mask along the image height.

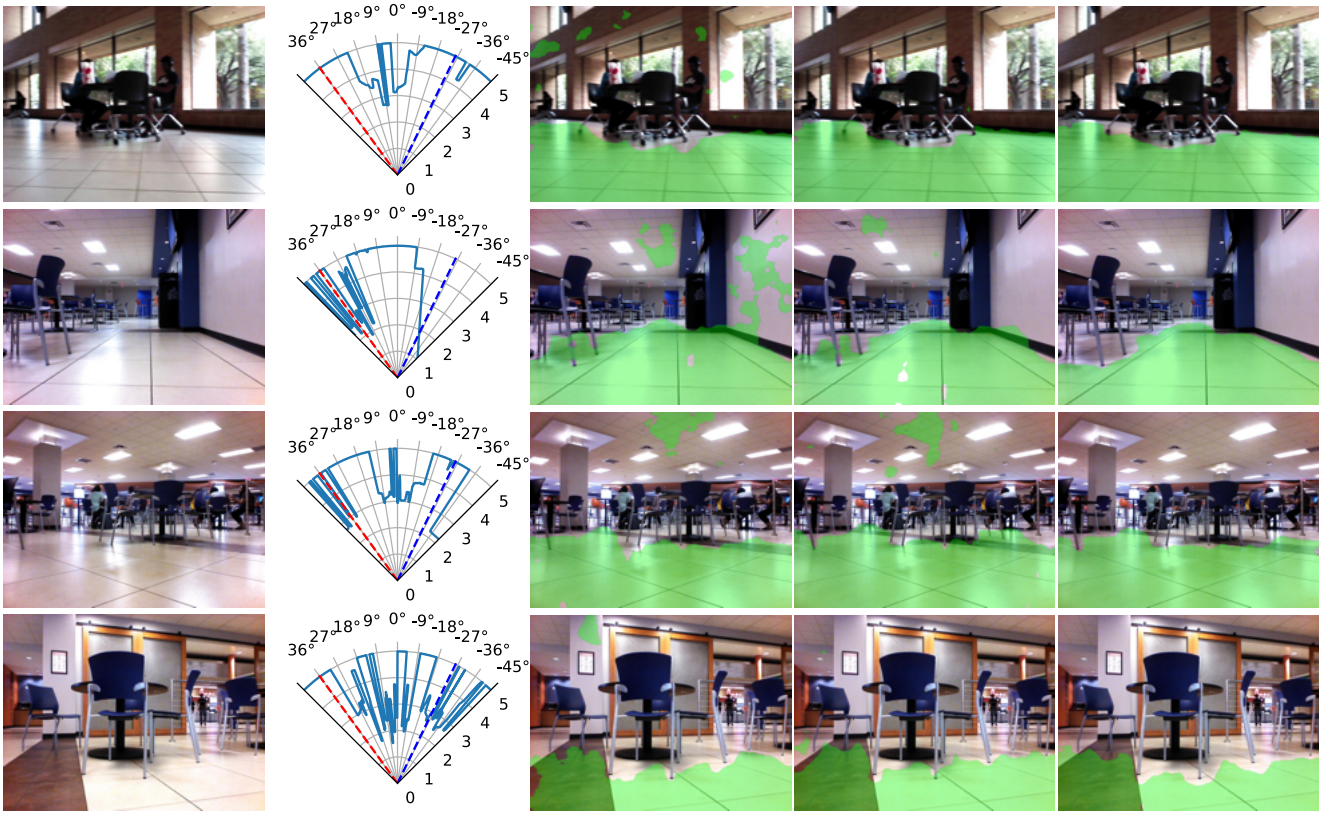

Fig. 5: Qualitative results on indoor traversability segmentation. Each row shows (1) the query RGB image, (2) the corresponding depth vector, (3) predictions without the two-stage depth attention module and without NCL, (4) predictions with the depth module but without NCL, and (5) predictions from the full model. The proposed depth module helps separate floors from walls/ceilings, while the NCL branch further improves recognition of thin obstacles (e.g., chair legs).

Finally, the full two-stage attention module (+H +W) delivers the best performance at 88.95 mIoU, a +11.5 improvement over the baseline. This demonstrates that horizontal and vertical attentions provide complementary benefits, and their joint modeling is crucial for capturing fine-grained geometric structures. While the height attention effectively propagates learned structural priors from RGB vertically, it can introduce hallucinations in regions with insufficient cues - as illustrated in our attached demo video.

Backbone	Setting	1-shot		
		Freespace	Obstacles	mIoU
DFormer [13]	+$p2p$ -$n2p$	88.36	72.96	80.66
	+$p2p$ +$n2p$	93.52	84.37	88.95

TABLE IV: Ablation study on the negative contrastive learning branch $n2p$, where $p2p$ represents the traditional positive prototype matching branch.

2) Negative Contrastive Learning Branch: To assess the contribution of the proposed negative contrastive learning (NCL) branch, we ablate the $p2p$ and $n2p$ branches under the 1-shot setting with the DFormer fusion backbone. Results are shown in Table IV.

Using only the positive prototype matching branch (+$p2p$ -$n2p$) achieves 80.66 mIoU. Adding the NCL branch (+$p2p$ +$n2p$) improves performance gain of +8.3. Notably, the largest improvement is observed in the obstacle class (+11.4 IoU), compared to +5.2 for freespace. This confirms that explicitly modeling negative prototypes significantly enhances the model's ability to separate obstacles from freespace—an aspect that purely positive prototype matching tends to overlook.

Therefore, the $n2p$ branch not only improves overall segmentation accuracy but also addresses a critical weakness of conventional FSS methods, i.e., the inability to robustly reject non-traversable regions.

V. Conclusion

We introduce a multi-modal RGB-D few-shot segmentation framework to improve indoor traversability analysis, with a particular focus on recognizing and excluding thin obstacles that are often overlooked by vision-only models. By digesting RGB images alongside 1D laser depth data, our approach leverages complementary geometric cues while remaining lightweight in terms of trainable parameters. To address the limited availability of labeled data and the need for generalization to unseen environments, we adopted a few-shot learning paradigm and proposed a novel negative contrastive learning branch to complement traditional positive prototype matching. Extensive experiments demonstrate

that our framework significantly improves both qualitative and quantitative performance, particularly in challenging scenarios with thin obstacles. We believe this work provides a promising direction for safer and more robust indoor robotic navigation, and we plan to release our dataset and code to further support research in the robotics and assistive systems community.

ACKNOWLEDGEMENTS

We thank Christos Sevastopoulos and Sneh Acharya for their invaluable assistance in collecting and annotating the dataset.

REFERENCES

[1] S. Hosseinpoor, J. Torresen, M. Mantelli, D. Pitto, M. Kolberg, R. Maffei, and E. Prestes, "Traversability analysis by semantic terrain segmentation for mobile robots," in *2021 IEEE 17th international conference on automation science and engineering (CASE)*. IEEE, 2021, pp. 1407–1413.

[2] C. Sevastopoulos and S. Konstantopoulos, "A survey of traversability estimation for mobile robots," *IEEE Access*, vol. 10, pp. 96 331–96 347, 2022.

[3] Q. An, C. Sevastopoulos, and F. Makedon, "Enhancing robustness of indoor robotic navigation with free-space segmentation models against adversarial attacks," 2024.

[4] Q. An, C. Sevastopoulos, F. Farahanipad, and F. Makedon, "Few-shot traversability segmentation of indoor robotic navigation with contrastive logits align," in *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2024, pp. 1959–1964.

[5] C. Sevastopoulos, J. Hussain, Q. An, S. Konstantopoulos, V. Karkaletsis, and F. Makedon, "Learning indoors free-space segmentation for a mobile robot from positive instances," in *2023 Seventh IEEE International Conference on Robotic Computing (IRC)*. IEEE, 2023, pp. 21–24.

[6] N. Hirose, A. Sadeghian, M. Vázquez, P. Goebel, and S. Savarese, "Gonet: A semi-supervised deep learning approach for traversability estimation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3044–3051.

[7] M. Oh, E. Jung, H. Lim, W. Song, S. Hu, E. M. Lee, J. Park, J. Kim, J. Lee, and H. Myung, "Travel: Traversable ground and above-ground object segmentation using graph representation of 3d lidar scans," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7255–7262, 2022.

[8] J. Watson, M. Firman, A. Monszpart, and G. J. Brostow, "Footprints and free space from a single color image," in *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[10] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[11] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*. Springer, 2012, pp. 746–760.

[12] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

[13] B. Yin, X. Zhang, Z. Li, L. Liu, M.-M. Cheng, and Q. Hou, "Dformer: Rethinking rgbd representation learning for semantic segmentation," *arXiv preprint arXiv:2309.09668*, 2023.

[14] H. Zhou, L. Qi, Z. Wan, H. Huang, and X. Yang, "Rgb-d co-attention network for semantic segmentation," in *Proceedings of the Asian conference on computer vision*, 2020.

[15] A. Casanova, P. O. Pinheiro, N. Rostamzadeh, and C. J. Pal, "Reinforced active learning for image segmentation," *arXiv preprint arXiv:2002.06583*, 2020.

[16] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8334–8343.

[17] Z. Lu, S. He, X. Zhu, L. Zhang, Y.-Z. Song, and T. Xiang, "Simpler is better: Few-shot semantic segmentation with classifier weight transformer," in *ICCV*, 2021.

[18] G. Zhang, S. Navasardyan, L. Chen, Y. Zhao, Y. Wei, H. Shi, *et al.*, "Mask matching transformer for few-shot segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 823–836, 2022.

[19] N. Hirose, A. Sadeghian, F. Xia, R. Martín-Martín, and S. Savarese, "Vunet: Dynamic scene view synthesis for traversability estimation using an rgb camera," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 2062–2069, 2019.

[20] C. Sevastopoulos, M. Theofanidis, M. Z. Zadeh, S. Acharya, S. Konstantopoulos, V. Karkaletsis, and F. Makedon, "Indoors traversability estimation with less labels for mobile robots," in *2022 Sixth IEEE International Conference on Robotic Computing (IRC)*. IEEE, 2022, pp. 306–311.

[21] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao, "Adversarial sensor attack on lidar-based perception in autonomous driving," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 2267–2281.

[22] Y. Zhu, C. Miao, F. Hajiaghajani, M. Huai, L. Su, and C. Qiao, "Adversarial attacks against lidar semantic segmentation in autonomous driving," in *Proceedings of the 19th ACM conference on embedded networked sensor systems*, 2021, pp. 329–342.

[23] K. Yang, L. M. Bergasa, E. Romera, and K. Wang, "Robustifying semantic cognition of traversability across wearable rgb-depth cameras," *Applied optics*, vol. 58, no. 12, pp. 3141–3155, 2019.

[24] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[25] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[26] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[27] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9197–9206.

[28] C. Lang, G. Cheng, B. Tu, and J. Han, "Learning what not to segment: A new perspective on few-shot segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8057–8067.

[29] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *IEEE Transactions on intelligent transportation systems*, vol. 24, no. 12, pp. 14 679–14 694, 2023.

[30] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[32] 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*, 2022, pp. 4248–4257.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[34] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.