SAB3R: Semantic-Augmented Backbone in 3D Reconstruction

Xuweiyi Chen[1]* Tian Xia[2]* Sihan Xu[2] Jianing Yang[2]
Joyce Chai[2] Zezhou Cheng[1]
[1] University of Virginia [2] University of Michigan

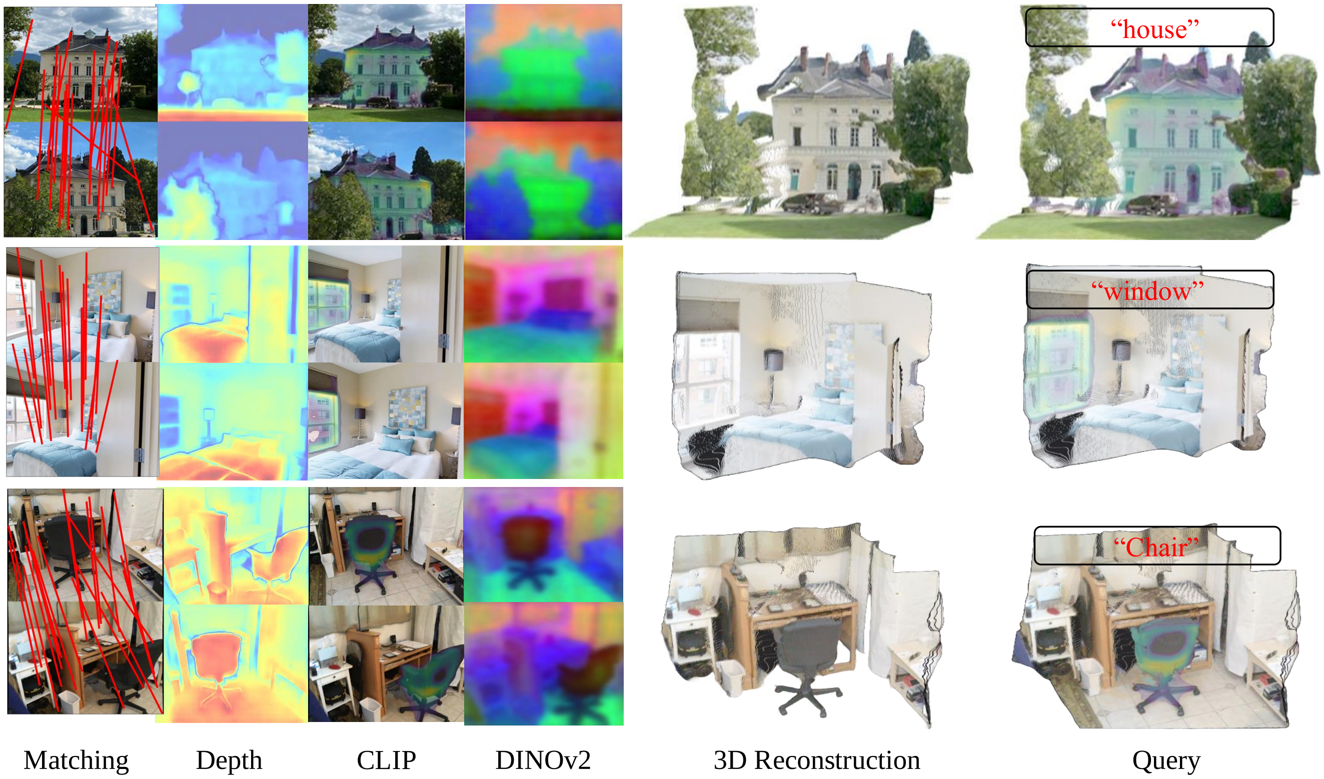

Figure 1. Our method, semantic-augmented backbone for 3D reconstruction, dubbed **SAB3R** , enables zero-shot open-vocabulary segmentation and 3D reconstruction from unposed images in a single forward pass. Building on a state-of-the-art 3D foundation model, SAB3R generates dense 2D foundational features, such as CLIP [57] and DINOv2 [52]. Additionally, SAB3R introduces a new capability by simultaneously performing 3D reconstruction and open-vocabulary semantic segmentation.

Abstract

The emergence of 3D vision foundation models (VFMs) represents a significant breakthrough in 3D computer vision. However, these models often lack robust semantic understanding due to the scarcity of 3D-language paired data. In contrast, 2D foundation models, trained on abundant data, excel in semantic tasks. In this work, we propose a novel distillation approach, ***SAB3R****, that transfers dense, per-pixel semantic features from 2D VFMs to enhance 3D VFMs. Our method achieves 2D semantic-aware feature integration while retaining the spatial reasoning capabilities of 3D VFMs.*

We validate our approach by showing that distillation does not compromise the base 3D foundation model, as demonstrated through evaluations on depth estimation and multi-view pose regression. Additionally, we introduce a new task, ***Map and Locate****, to showcase the novel capability of Multi-view 3D Open Vocabulary Semantic Segmentation. Finally, the experiment of our method reveals its ability to maintain a robust understanding of 3D structures while markedly improving its 2D semantic comprehension. Our results highlight the effectiveness of our approach.*

[*] Equal contribution

1. Introduction

2D Vision Foundation Models (VFMs), such as CLIP [57], SAM [34], and DINOv2 [52], have emerged as versatile backbones for a wide range of vision tasks. By leveraging abundant 2D data, these models demonstrate strong semantic understanding, excelling in applications like few-shot segmentation [2] and label-free scene understanding [9]. Their role as general-purpose feature extractors enhances performance in downstream tasks, such as segmentation, depth estimation, and correspondence [15, 80, 81].

In parallel, 3D foundation models like DUSt3R [79] and MASt3R [39] have advanced 3D scene reconstruction from unposed image collections. These models integrate tasks such as camera calibration, depth estimation, correspondence, and pose estimation into a single forward pass, showcasing exceptional multiview reasoning and geometric understanding.

While 2D VFMs excel in semantic comprehension and capturing features like depth and surface normals [15], they lack multiview consistency, especially under significant viewpoint variation. Conversely, 3D models, such as those built on CroCo [83] and CroCoV2 [82], are highly effective for multiview and geometric tasks but exhibit weaker 2D semantic understanding. These complementary strengths highlight an opportunity to combine 2D and 3D approaches, addressing their respective limitations to drive progress in visual understanding.

Humans, by contrast, seamlessly interpret images by combining 2D visual information with an intuitive understanding of 3D structure, a skill refined through lifelong experience with depth and motion [36]. This drives the need for models that can integrate both semantic understanding and 3D understanding. Moreover, maintaining separate models for various vision tasks is inefficient due to high memory and runtime costs, particularly on edge devices, and it forgoes the advantages of cross-model learning [64].

While multitask learning [10] offers a promising approach to addressing these inefficiencies by enabling shared learning across multiple tasks, its application to 3D vision remains hindered by a fundamental challenge: the lack of large-scale, high-quality 3D-language paired datasets. In 2D vision, the success of state-of-the-art models such as CLIP [57] can be attributed to their training on massive datasets like LAION [67] and DataComp [23], each containing billions of samples. These datasets provide the diversity and scale necessary for training robust multimodal models. However, the 3D domain lags far behind; even the largest 3D-language datasets, such as SceneVerse [31] and 3D-Grand [88], are an order of magnitude smaller. This stark contrast underscores a significant data gap between the 2D vision and 3D vision.

To address the above challenges, we begin by posing two key questions:

- *Can we distill 2D foundation models into 3D foundation models without adding parameters, thereby enhancing 2D semantic understanding while preserving 3D capabilities?*
- *Can co-training with 3D and semantic tasks optimize semantic understanding?*

We propose SAB3R , as illustrated in Fig. 2, to tackle these challenges. Our approach distills 2D foundation models into 3D foundation models without introducing additional parameters, enhancing 2D semantic understanding while retaining 3D capabilities. Building on prior work [77], we identify feature resolution as critical to effective distillation. To address this, we employ FeatUp [22] to upsample features to any desired resolution, ensuring better alignment and integration of 2D and 3D representations.

During training, we initialize SAB3R with MASt3R's original weights, retaining its tasks and objectives. Using FeatUp, we extract dense 2D features from MaskCLIP [13] and DINOv2 [52], which are regressed with two added DPT heads. After training, SAB3R generates depth, dense CLIP features, and dense DINO features simultaneously.

Our model, SAB3R , introduces a novel capability called *Map and Locate*, enabling multi-view 3D open vocab semantic segmentation. To evaluate this, we present a new benchmark where SAB3R significantly outperforms a pipelined baseline (FeatUp + MASt3R). In addition, experiments show that SAB3R achieves comparable performance to MASt3R on 3D tasks while matching FeatUp's MaskCLIP on semantic tasks.

In summary, our contributions are:

- **SAB3R** : A unified model that integrates dense 2D semantic understanding into a 3D foundation model, preserving 3D capabilities while enhancing 2D semantics through a parameter-efficient distillation process.
- ***Map and Locate* Benchmark**: A new task for evaluating multi-view 3D semantic segmentation.

2. Related Work

2.1. Knowledge Distillation

Bucila et al. [5] and Hinton et al. [27] first introduced the concept of knowledge distillation, where a smaller, compressed model (student) is trained using knowledge from a larger, pretrained model (teacher). Building on this foundation, recent studies have applied distillation to vision-language models (VLMs), such as EVA [18, 19], DIMEFM [74], CLIPPING [55], and CLIP-KD [87], achieving transfer of the teacher's zero-shot capabilities to the student model. Recently, AM-RADIO [60] aggregate multi vision foundation model into one model, which demonstrate comprehensive knowledge distillation setups and tricks. Our work diverges significantly in that we focus on developing a 3D foundation model, aiming to establish

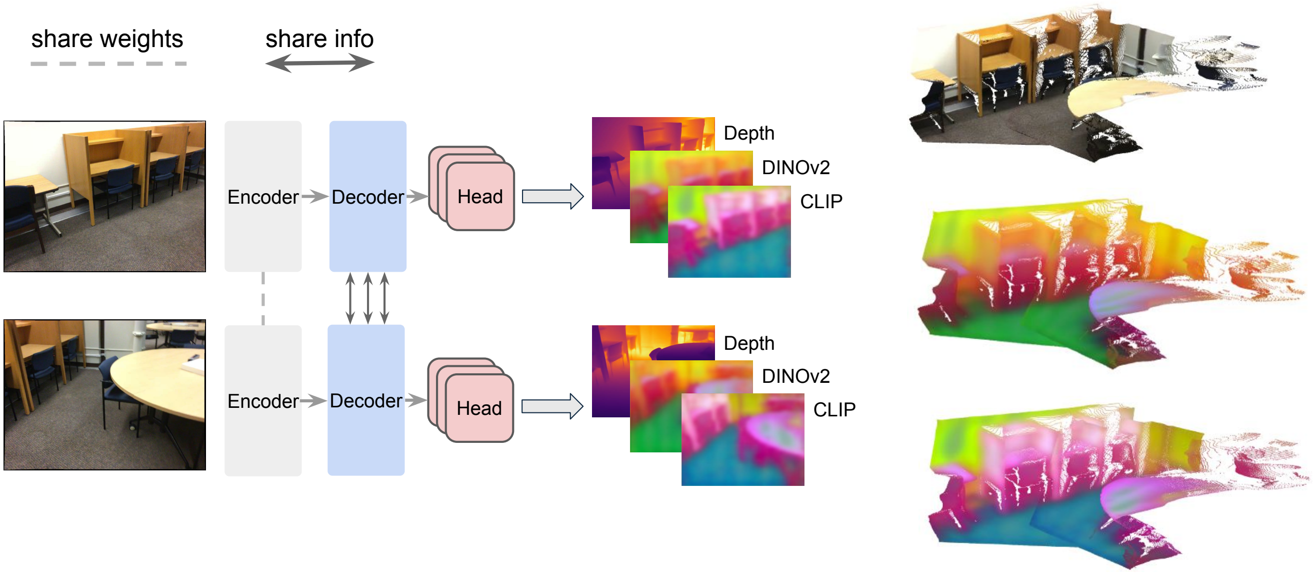

Figure 2. **Methods Architecture.** We distill dense features from CLIP and DINO into the MASt3R framework, enriching it with 2D semantic understanding. Each encoder-decoder pair operates on multi-view images, sharing weights and exchanging information to ensure consistent feature extraction across views. The model simultaneously generates depth, dense DINOv2, and dense CLIP features, which are then used for multi-view 3D reconstruction and semantic segmentation. This architecture enables SAB3R to seamlessly integrate 2D and 3D representations, achieving both geometric and semantic comprehension in a unified model.

a model proficient in executing both 3D and 2D tasks at the same time.

Our setup is also closely related to Continual Learning [54], where models incrementally learn new information. A key challenge in this area is catastrophic forgetting [46, 47], where newly acquired knowledge overwrites previously learned information. Continual Learning approaches combat forgetting through strategies like regularization [35, 95], experience replay [26, 61], regularized replay [8, 20], dynamic expansion [68, 86], and optimization-based methods [49, 53].

In this work, we find that our model SAB3R does not show severe sign of catastrophic forgetting. Our pipeline supports the model to adapt to new capabilities without losing previously learned knowledge.

2.2. 3D Language Grounding

Most 3D visual-language models operate directly on 3D point clouds without leveraging 2D pre-trained features. SAT-2D [89] was one of the first 3D visual grounding models to incorporate 2D visual features, aligning 2D and 3D representations during training and achieving significant improvements over versions without 2D features. More recent approaches, such as 3DLLM [28] in 3D Question Answering, use multi-view 2D features with LLMs to decode answers, but have yet to fully address 3D visual grounding tasks. Similarly, PQ3D [102] integrates various visual backbones, including a 2D feature backbone from OpenScene [56].

EFM3D [72] lifts 2D image features into 3D feature volumes, but focuses on 3D object detection and surface reconstruction. ODIN [30] proposes an interleaved 2D-3D backbone with pre-trained 2D weights, but is limited to object detection. Fit3D [93], which lifts 2D semantic features into 3D Gaussian representations, injects 3D awareness when training 2D foundation models—a complementary approach to ours.

2.3. Distilling 2D Features into 3D

Our work is closely related to research focused on distilling 2D features into 3D representations. Previous approaches often leverage novel view synthesis methods, such as NeRF [48] and Gaussian Splatting [32], which aggregate information across multiple viewpoints. For example, semantic NeRF [99] and Panoptic Lifting [70] embed semantic information into 3D, while methods like LeRF [33], Distilled Feature Fields [69], NeRF SOS [17], and Neural Feature Fusion Fields [76] incorporate pixel-aligned feature vectors from models such as Lseg [40] and DINO [6]. Recently, Featured 3DGS [101] has distilled 2D pre-trained models into Gaussian splatting for enhanced 3D feature representation.

While our approach shares this goal of integrating 2D imagery for 3D reconstruction and feature embedding, it offers greater flexibility. Unlike prior work, we do not require posed images and achieve a higher level of generalization, avoiding the need for scene-specific tuning.

3. Method

In this section, we describe our approach for distilling dense 2D Vision Foundation Model (VFM) features into a 3D VFM. Starting from a base 3D VFM, we transfer knowledge from 2D VFM features enhanced with FeatUp. Our goal is to integrate 2D and 3D knowledge into a unified backbone, enabling simultaneous 3D reconstruction and open-vocabulary semantic segmentation.

To clarify our method, this section is structured as follows: Sec. 3.1 provides a summary of the foundational components, Sec. 3.2 describes the distillation of 2D semantic features into the model, and Sec. 3.3 explains how additional features can be incorporated to enhance model capabilities.

3.1. Foundational Components

DUSt3R [79] is a recent method that addresses a range of 3D tasks using unposed images as input, including camera calibration, depth estimation, pixel correspondence, camera pose estimation, and dense 3D reconstruction. It uses a transformer-based network to generate *local* 3D reconstructions from two input images, producing dense 3D point clouds $X^{1,1}$ and $X^{2,1}$, referred to as *pointmaps*.

A pointmap $X^{a,b} \in \mathbb{R}^{H\times W\times 3}$ represents a 2D-to-3D mapping from each pixel $i = (u, v)$ in image I^a to its corresponding 3D point $X^{a,b}_{u,v} \in \mathbb{R}^3$ in the coordinate system of camera C^b. By jointly regressing two pointmaps, $X^{1,1}$ and $X^{2,1}$, expressed in the coordinate system of camera C^1, DUSt3R simultaneously performs calibration and 3D reconstruction. For multiple images, a global alignment step merges all pointmaps into a unified coordinate system.

Images are encoded in a Siamese manner using a ViT [14], producing representations H^1 and H^2:

$$H^1 = \text{Encoder}(I^1), \quad H^2 = \text{Encoder}(I^2).$$

Two intertwined decoders process these representations, exchanging information via cross-attention to capture spatial relationships and global 3D geometry. The enhanced representations are denoted H'^1 and H'^2:

$$H'^1, H'^2 = \text{Decoder}(H^1, H^2).$$

Finally, prediction heads regress the pointmaps and confidence maps:

$$X^{1,1}, C^1 = \text{Head}^1_{3D}([H^1, H'^1]), \tag{1}$$

$$X^{2,1}, C^2 = \text{Head}^2_{3D}([H^2, H'^2]). \tag{2}$$

3.2. Distilling 2D Semantic Features

To integrate 2D semantic information into the model while retaining its 3D capabilities, we design a multitask framework that prevents catastrophic forgetting. This framework enables the model to simultaneously learn both 2D and 3D features. We adopt the MASt3R [39] architecture, which consists of a ViT-Large encoder, a ViT-Base decoder, and DPT heads. To distill dense 2D features, we introduce new heads to regress features from DINO [52] and CLIP [57].

Following DUSt3R [79] and MASt3R [38], the new heads leverage either a DPT architecture or a simpler MLP structure. The DPT design is particularly effective for dense prediction tasks like depth estimation and semantic feature extraction. In addition to the depth and descriptor heads ($\text{Head}^{1,2}_{3D}$ and $\text{Head}^{1,2}_{desc}$), we introduce two new heads, $\text{Head}^{1,2}_{\text{2D feature}}$, for distilling 2D features:

$$S^1 = \text{Head}^1_{\text{2D feature}}([H^1, H'^1]), \tag{3}$$

$$S^2 = \text{Head}^2_{\text{2D feature}}([H^2, H'^2]). \tag{4}$$

Here, H^1 and H^2 are embeddings from the encoder, and H'^1, H'^2 are enhanced representations from the decoder. The concatenation $[H, H']$ combines spatial and semantic information.

To preserve depth estimation capabilities, we retain the regression loss $\mathcal{L}_{conf}$ from DUSt3R and the matching loss $\mathcal{L}_{match}$ from MASt3R. Additionally, we introduce a regression loss for the 2D features, guiding the model to learn semantic information:

$$\mathcal{L}_{\text{2D}} = \left\| S^v - \hat{S}^v \right\|, \quad v \in \{1, 2\}, \tag{5}$$

where $\hat{S}^v$ is the target 2D feature extracted by MaskCLIP or DINOv2 for the corresponding view v. Dense pixel features from FeatUp are used as supervision.

The total loss combines all components, weighted by hyperparameters β and γ:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{conf} + \beta\mathcal{L}_{match} + \gamma\mathcal{L}_{\text{2D}}. \tag{6}$$

3.3. Incorporating Additional Features

Our distillation pipeline is designed to flexibly incorporate multiple 2D features into the 3D foundation model, enhancing its capabilities. For each additional feature, we add a dedicated head and regression loss, resulting in an updated training objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{conf} + \beta\mathcal{L}_{match} + \gamma_1\mathcal{L}_{2D_1} + \gamma_2\mathcal{L}_{2D_2}. \tag{7}$$

Here, $\mathcal{L}_{2D_1}$ and $\mathcal{L}_{2D_2}$ are regression losses for individual 2D features, with γ_1 and γ_2 controlling their contributions. MaskCLIP and DINOv2 features are integrated into the 3D backbone through this framework, with dedicated heads for each feature.

4. A Novel Task: *Map and Locate*

Task Setting In this novel task, termed *Map and Locate*, the model receives multiview inputs and a set of semantic labels to reconstruct a 3D scene. This task extends beyond independent depth estimation for each image, requiring the model to infer relative camera poses across views and classify the semantic category of each predicted 3D point.

The task is defined as follows: given n input images ($n \geq 2$) and a predefined set of L semantic classes $\mathcal{L} = \{0, \ldots, L-1\}$, the goal is to map each pixel i to a pair $(X_i, l_i) \in \mathbb{R}^3 \times \mathcal{L}$, where $X_i = (x_i, y_i, z_i)$ represents the 3D coordinates of the point corresponding to pixel i, and l_i denotes its semantic class. For an image I of resolution $W \times H$, this establishes a one-to-one mapping between pixels and 3D scene points with semantic labels, i.e., $I_{i,j} \leftrightarrow (X_{i,j}, l_{i,j})$, for all $(i,j) \in \{1, \ldots, W\} \times \{1, \ldots, H\}$. We assume each camera ray intersects only a single 3D point, excluding cases like translucent surfaces. Ambiguous or out-of-class pixels are assigned a void label in the annotations.

For implementation, we adopt MaskCLIP [13] enhanced with FeatUp [22], combined with the MASt3R [39] pipeline as our baseline method. MaskCLIP and MASt3R act as teacher models for our approach, guiding the distillation process to achieve both 3D reconstruction and open-vocabulary semantic segmentation.

Data Curation Our data is sourced from ScanNet [11], a large-scale indoor scene dataset that provides RGB-D sequences, camera poses, and semantic and instance-level annotations. From the validation split, we curate a subset of 10 diverse scenes, selected for their unique object layouts and camera trajectories. For each scene, we create image groups containing 2, 3, or 4 views, ensuring that each image overlaps with at least one other in the group. This overlap guarantees shared visual context, enabling robust evaluation of 3D reconstruction and localization tasks. To balance evaluation time and dataset diversity, we limit our selection to 10 scenes, which already requires approximately 2 hours for the evaluation to complete.

For semantic classification, we map ground-truth annotations to the widely used NYU40 class taxonomy [50]. The curated dataset includes a total of 436 objects with both semantic and instance-level annotations. Each image group is paired with its corresponding RGB images, depth maps, camera poses (intrinsics and extrinsics), and semantic and instance labels. Detailed data statistics, example image groups, and the full data curation process, including selection criteria and preprocessing steps, are provided in the supplementary materials.

Evaluation Metrics For the *Map and Locate* task, we evaluate model performance using several key metrics, and in all metrics, higher values consistently indicate better performance. Additionally, before evaluating these metrics, models are required to compute pair (X, l) for every pixel in each image, using only the image inputs without any ground truth data, such as intrinsic or extrinsic matrices, then use one ground truth image's depth and pose for scaling and alignment to the ground truth coordinates.

Mean Intersection over Union (mIoU) quantifies the overlap between predicted and ground truth points, calculated as the ratio of correctly predicted points to the union of predicted and ground truth points. This metric provides an overall measure of segmentation accuracy. In our task, we compute the mIoU by finding the nearest predicted point for each ground truth point and using its label to evaluate against the ground truth labels.

Accuracy (Acc.) is defined as the proportion of correctly predicted points relative to the total ground truth points, indicating the model's effectiveness in assigning correct semantic classes to 3D points. In our setteing, Similar to mIoU, we calculate Acc using the same approach.

Mean Completeness (Comp.) measures how comprehensively the predicted points cover the ground truth point cloud. After aligning the predicted points with the ground truth pose, we compute the average distance from each predicted point to its nearest neighbor in the ground truth, offering a general sense of the reconstruction's completeness. For our task, we filter points based on each test label in both the ground truth and the predictions, then calculate the Comp. metric accordingly.

Median Completeness (Median Comp.) is similar to mean completeness but calculates the median of nearest-neighbor distances instead. This approach reduces the impact of outliers, providing a more stable indication of coverage consistency across samples.

5. Experiments

In this section, we showcase the effectiveness of our approach for distilling 2D foundation models into a 3D foundation model. The section is organized into five parts. In Sec.5.1, we provide details of the implementation. Sec.5.2 analyzes how SAB3R retains performance without exhibiting catastrophic forgetting, compared to the teacher models. In Sec.5.3, we demonstrate our method's zero-shot semantic segmentation performance, achieving results comparable to the teacher models. Sec.5.4 presents results and analysis for the novel task, *Map and Locate*. Finally, in Sec. 5.5, we provide evidence that the encoder not only preserves 3D capabilities but also enhances 2D semantic understanding.

5.1. Implementation Details

We fine-tune our model using datasets from DUSt3R and MASt3R, including Habitat [75], ScanNet++ [91], ARKitScenes [1], Co3Dv2 [62], and BlenderMVS [90]. Data preprocessing adheres to the guidelines of each dataset. To avoid the impracticality of storing dense 2D VFM features locally, which would require over 60 TB of storage, we

Methods	Train	NYUD-v2 (Indoor)		KITTI (Outdoor)	
		Rel↓	$\delta_{1.25}$ ↑	Rel↓	$\delta_{1.25}$ ↑
DPT-BEiT[59]	D	**5.40**	**96.54**	9.45	89.27
NeWCRFs[92]	D	6.22	95.58	**5.43**	**91.54**
Monodepth2 [25]	SS	16.19	74.50	11.42	86.90
SC-SfM-Learners [4]	SS	13.79	79.57	11.83	86.61
SC-DepthV3 [73]	SS	**12.34**	**84.80**	11.79	86.39
MonoViT [98]	SS	-	-	**9.92**	**90.01**
RobustMIX [51]	T	11.77	90.45	18.25	76.95
SlowTv [71]	T	11.59	87.23	(6.84)	(56.17)
DUSt3R 224-NoCroCo	T	14.51	81.06	20.10	71.21
DUSt3R 224	T	10.28	88.92	16.97	77.89
DUSt3R 512	T	**6.51**	**94.09**	**12.02**	**83.43**
MASt3R	T	8.17	92.59	**8.28**	**93.27**
SAB3R (C)	T	7.80	92.67	11.63	86.74
SAB3R (CD)	T	**7.67**	**92.82**	12.53	83.51

Table 1. **Monocular depth estimation on NYU-v2 and KITTI datasets.** D = Supervised, SS = Self-supervised, T = Transfer (zero-shot). (Parentheses) refers to training on the same set. SAB3R (C) represents our model distilled with CLIP features, while SAB3R (CD) builds upon this by integrating both CLIP and DINO features during distillation. This notation is used consistently throughout the paper.

Methods	RRA@15↑	RTA@15↑	mAA(30)↑
Colmap+SG [12, 65]	36.1	27.3	25.3
PixSfM [43]	33.7	32.9	30.1
RelPose [96]	57.1	-	-
PosReg [78]	53.2	49.1	45.0
PoseDiff [78]	80.5	79.8	66.5
RelPose++ [42]	(85.5)	-	-
RayDiff [97]	(93.3)	-	-
DUSt3R-GA [79]	**96.2**	86.8	76.7
DUSt3R [79]	94.3	88.4	77.2
MASt3R	94.15	**88.58**	**81.14**
SAB3R (C)	92.57	87.31	79.66
SAB3R (CD)	92.93	87.76	80.25

Table 2. **Multi-view pose regression on the CO3Dv2 [62] dataset using 10 random frames.** Results in parentheses denote methods evaluated on 8 views, as they do not report results for the 10-view setup. We distinguish multi-view and pairwise methods for clarity.

leverage FeatUp to dynamically generate these features during training. Additional details on the datasets and preprocessing steps are provided in the supplementary materials.

Training We adopt MASt3R [39] as our base 3D foundation model. During training, we unfreeze the encoder to enhance its ability to produce 2D semantic-aware features while maintaining depth accuracy. For distilling only MaskCLIP features, we set $\beta = 0.75$ and $\gamma = 20$. When distilling both MaskCLIP and DINOv2 features, we adjust these weights to $\beta = 0.75$, $\gamma_1 = 20$, and $\gamma_2 = 4$.

Computational Resources Each checkpoint is optimized around 3 days, using either 8 A40 GPUs or 4 A100 80GB GPUs.

5.2. Zero-Shot 3D Tasks

Monocular Depth Estimation We benchmark SAB3R on both an indoor dataset, NYUv2 [50], and an outdoor dataset, KITTI [24], comparing its performance to state-of-the-art methods in Tab. 1. For monocular depth evaluation, we use two commonly applied metrics following DUSt3R [79] and recent studies [3, 71].

As shown in Tab. 1, SAB3R demonstrates strong adaptability to both indoor and outdoor environments. Distilling dense features from MaskCLIP or DINOv2 into the MASt3R backbone does not degrade the model's performance or induce catastrophic forgetting. Therefore, SAB3R is still capable of making accurate depth prediction. Notably, while MASt3R's performance decreases relative to its DUSt3R backbone, our training approach preserves robust monocular depth estimation capabilities. Interestingly, SAB3R trained with MaskCLIP, or with both MaskCLIP and DINOv2, outperforms the base model MASt3R on the NYUv2 indoor dataset [50]. However, our approach performs less effectively in outdoor scenarios, likely due to the indoor-focused nature of our training data.

Relative Camera Pose Next, we evaluate for the task of relative pose estimation on the CO3Dv2 [62] dataset. CO3Dv2 contains 6 million frames extracted from approximately 37k videos, covering 51 MS-COCO categories.

We compare our method's Relative Camera Pose results with popular approaches like RelPose [96], RelPose++ [42], PoseReg and PoseDiff [78], RayDiff [97], DUSt3R [79] and MASt3R [39] in Tab. 2. Our experiments show that our method performs comparably to the original MASt3R [39], indicating that catastrophic forgetting is not an issue. These results reinforce that SAB3R retains strong relative camera pose capabilities and can reliably estimate camera poses from unposed images.

5.3. Zero-Shot Open Vocabulary Tasks

Zero-Shot Transfer to Semantic Segmentation We evaluate the semantic features learned by SAB3R through zero-shot semantic segmentation on two datasets: Pascal VOC [16] and ADE20K [100]. The results, shown in Table 4, follow the evaluation protocol of SAM-CLIP [77], with the key difference being that SAB3R outputs dense pixel-level predictions. Notably, SAB3R outperforms SAM-CLIP on the challenging ADE20K benchmark, which contains 150 semantic classes. While SAB3R does not surpass SAM-CLIP on the Pascal VOC dataset, it delivers com-

Model	Sparse View = 2					Sparse View = 3					Sparse View = 4				
	mIoU	**Acc.**	**Comp.**	**Median Comp.**	**Time (s)**	**mIoU**	**Acc.**	**Comp.**	**Median Comp.**	**Time (s)**	**mIoU**	**Acc.**	**Comp.**	**Median Comp.**	**Time (s)**
Baseline	4.57	18.10	0.64	0.67	3.92	6.03	21.26	0.68	0.71	22.74	5.12	19.31	0.68	0.70	36.72
SAB3R (C)	17.26	41.11	**0.73**	0.75	1.92	22.83	**53.19**	**0.78**	**0.81**	7.49	19.92	**48.07**	**0.77**	**0.80**	9.97
SAB3R (CD)	**17.50**	**42.72**	**0.73**	**0.76**	2.54	**22.94**	52.86	0.77	0.80	8.67	**20.31**	46.26	0.75	0.78	12.15

Table 3. Performance comparison across different sparse view configurations (2, 3, and 4 views) with metrics including mIoU, Accuracy, Mean Completeness, Median Completeness, and Inference Time. The Inference Time refers to both reconstruction time and CLIP feature extraction.

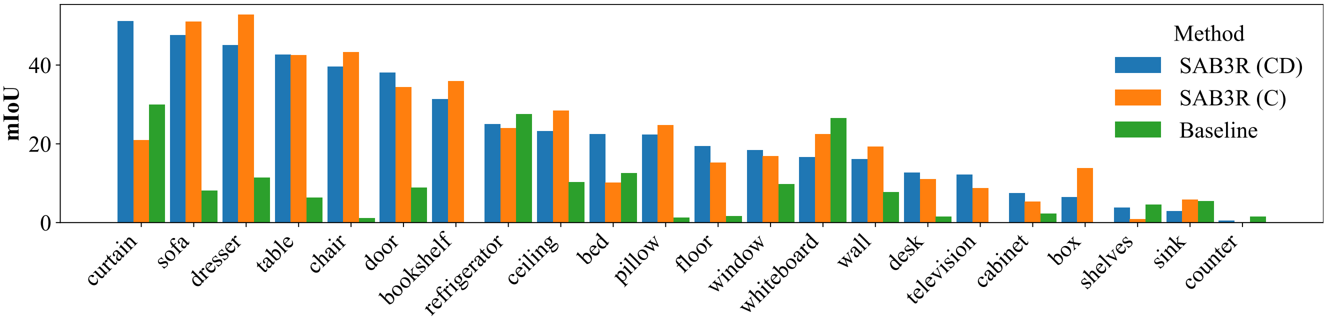

Figure 3. **mIoU Analysis on Frequently Occurring Objects Across Three Methods (Sparse View = 3).** This plot compares mIoU values for frequently appearing objects, illustrating performance differences between our methods and the pipeline approaches and providing insights into the superior results achieved by our methods.

Model	Arch	VOC↑	ADE20k↑
GroupViT [84]	ViT-S	52.3	-
ViewCo [63]	ViT-S	52.4	-
ViL-Seg [44]	ViT-B	37.3	-
OVS [85]	ViT-B	53.8	-
CLIPpy [58]	ViT-B	52.2	13.5
TCL [7]	ViT-B	51.2	14.9
SegCLIP [45]	ViT-B	52.6	8.7
SAM-CLIP [77]	ViT-B	**60.6**	17.1
FeatUp (MaskCLIP)	-	51.2	14.29
SAB3R (C)	ViT-B	55.4	18.29
SAB3R (CD)	ViT-B	56.4	**19.04**

Table 4. **Zero-shot Semantic Segmentation Comparison.** Performance comparison of zero-shot semantic segmentation with recent state-of-the-art methods. **Note:** Results for SAB3R are based solely on the CLIP-head output.

petitive performance and outperforms the teacher model, FeatUp-upsampled MaskCLIP [13].

Our results suggest that 3D-aware semantic features learned by SAB3R are particularly effective on large surface areas such as walls and floors. Furthermore, qualitative examples, including PCA visualizations of the jointly learned 2D features, are provided in the supplementary material.

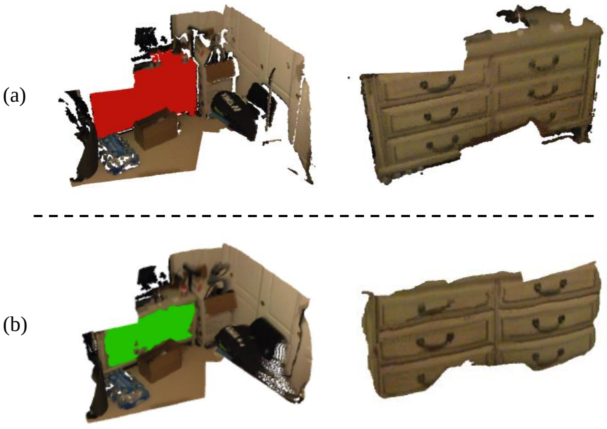

Figure 4. **Qualitative Example of *Map and Locate*.** This figure illustrates an example from our benchmark. In (a), the ground truth annotation for the scene is highlighted in red, with the dresser segmented from the rest of the scene on the left. In (b), the predictions from SAB3R are highlighted in green, and the predicted dresser is similarly segmented on the right. These segmented results are subsequently used to compute evaluation metrics.

5.4. A Novel Task - *Map and Locate*

We present the results in Tab. 3. Our method, SAB3R , outperforms the baseline model by a significant margin. Moreover, our inference speed is 3 times faster than the baseline, as SAB3R is an end-to-end model, whereas the pipeline-based baseline requires inference from two separate mod-

Table 5. Head feature probing evaluations on classification, text-to-image retrieval, and correspondence tasks, comparing SAB3R with state-of-the-art models using the ViT-B architecture. SAB3R demonstrates minimal forgetting compared to baseline FMs on their original tasks.

Model	**IM1K (FT)**	**G. Corres.**	**S. Corres.**
MaskCLIP [13]	83.6	43.05	7.04
MASt3R [38]	48.5	66.2	11.9
SAB3R (C)	58.6	69.2	18.6
SAB3R (CD)	52.9	62.2	14.4

els. Additionally, we observe that our model consistently outperforms the baseline across all sparse views = 2, 3, 4 in every metric, demonstrating strong performance in the *Map and Locate* task. Specifically, in terms of mIoU and accuracy—metrics that emphasize semantic quality—our model outperforms the baseline by large margin in mIoU and accuracy, which highlights that, despite the challenges in pose free 3D reconstruction. Our method, SAB3R , can seamlessly do 3D reconstruction and semantic segmentation at the same time. For completion metrics, our model also surpasses the baseline in reconstructing semantic-labeled objects across all sparse view configurations. Additionally, we observe no direct correlation between the number of views and the resulting scores. We hypothesize that this is because an increase in views can enhance scores when multiple views concentrate on a specific object or region, enabling the model to better understand and reconstruct those parts. Conversely, scores may decrease as the number of views increases if the scene becomes more dispersed with minimal overlap between views, complicating the reconstruction process.

In Fig. 3, our model demonstrates significant improvements over the baseline in large furniture categories such as sofas, dressers, tables, and chairs. It also successfully recognizes items like bookshelves and televisions, which the baseline fails to detect. Across most categories, our model achieves substantially higher scores, showcasing its strong semantic understanding and superior 3D reconstruction capabilities. Furthermore, it exhibits the ability to identify smaller objects and less common items, underscoring its versatility and robustness.

In Fig. 4, we showcase an example of mapping and locating a *dresser* across two images. In part (b) of the qualitative example, the predicted segmentation demonstrates remarkable accuracy compared to the ground truth shown in part (a), highlighting the effectiveness of our model SAB3R .

5.5. Encoder-Probing Evaluation on Learned Representations

By integrating MASt3R [39] and MaskCLIP [13], we hypothesis that our model will inherit the representational strengths of its parent models. MASt3R [39] excels in 3D tasks, while MaskCLIP is adept at capturing high-level semantic visual information across entire images. We hope that our model SAB3R synergistically combines these strengths, enhancing its versatility across a wide range of downstream vision tasks. Table 5 highlights the evaluation results.

First, we present the ImageNet-1k probing experiment results for MaskCLIP [13], MASt3R, SAB3R distilled with CLIP, and SAB3R distilled with CLIP [57] and DINOv2 [52]. Our findings indicate that the encoder of SAB3R produce feature contain richer semantic information compared to the encoder of MASt3R [39].

Additionally, we provide accuracy results for geometric correspondence and semantic correspondence, SAB3R achieves performance comparable to MASt3R [39] on geometric correspondence. SAB3R achieves better performance on semantic correspondence compare with teacher models. This highlights the encoder clearly understand more 2D semantic while preserve strong 3D capabilities. Detailed descriptions of the metrics and probing evaluation methodology are included in the supplementary materials.

6. Conclusion

Our approach demonstrates that distillation preserves the fundamental capabilities of the base 3D Visual Foundation Model (VFM), as evidenced by strong performance in depth estimation and multi-view pose regression tasks. By introducing the novel **Map and Locate** task, we highlight the ability of our method to perform Multi-view 3D Open Vocabulary Semantic Segmentation effectively. This task showcases a unique combination of 3D structural understanding and 2D semantic reasoning, enabling precise segmentation across multiple views.

Additionally, our experiments reveal that our distillation process not only retains a robust understanding of 3D spatial structures but also significantly enhances the model's 2D semantic comprehension. This dual capability underscores the versatility and effectiveness of our approach. By bridging the strengths of 2D and 3D VFMs, our method creates a pathway for future advancements in combining 2D and 3D representations. As the field of VFMs continues to grow, we hope this work inspires the research community to explore new methods for integrating the complementary strengths of 2D and 3D models for a broader range of tasks.

7. Acknowledgement

The authors gratefully acknowledge Research Computing at the University of Virginia for providing the computational resources and technical support that made the results in this work possible. Research Computing at UVA.

References

[1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 5, 1

[2] Reda Bensaid, Vincent Gripon, François Leduc-Primeau, Lukas Mauch, Ghouthi Boukli Hacene, and Fabien Cardinaux. A novel benchmark for few-shot semantic segmentation in the era of foundation models, 2024. 2

[3] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Tat-Jun Chin, Chunhua Shen, and Ian Reid. Auto-rectify network for unsupervised indoor depth estimation, 2021. 6

[4] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision (IJCV)*, 2021. 6

[5] Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Knowledge Discovery and Data Mining*, 2006. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 3

[7] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023. 7

[8] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem, 2019. 3

[9] Runnan Chen, Youquan Liu, Lingdong Kong, Nenglun Chen, Xinge Zhu, Yuexin Ma, Tongliang Liu, and Wenping Wang. Towards label-free scene understanding by vision foundation models, 2023. 2

[10] Michael Crawshaw. Multi-task learning with deep neural networks: A survey, 2020. 2

[11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 5, 2

[12] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised Interest Point Detection and Description. In *CVPR*, 2018. 6

[13] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Maskclip: Masked self-distillation advances contrastive language-image pretraining, 2023. 2, 5, 7, 8, 1

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4

[15] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D Awareness of Visual Foundation Models. In *CVPR*, 2024. 2

[16] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 6

[17] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes, 2022. 3

[18] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale, 2022. 2

[19] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149: 105171, 2024. 2

[20] Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning, 2019. 3

[21] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution, 2024. 1

[22] Stephanie Fu, Mark Hamilton, Laura E. Brandt, Axel Feldmann, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 5, 1

[23] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023. 2

[24] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 6, 1

[25] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. 2019. 6

[26] Tyler L. Hayes, Nathan D. Cahill, and Christopher Kanan. Memory efficient experience replay for streaming learning, 2019. 3

[27] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 2

[28] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models, 2023. 3

[29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. 1

[30] Ayush Jain, Pushkal Katara, Nikolaos Gkanatsios, Adam W. Harley, Gabriel Sarch, Kriti Aggarwal, Vishrav Chaudhary, and Katerina Fragkiadaki. Odin: A single model for 2d and 3d segmentation, 2024. 3

[31] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. *arXiv preprint arXiv:2401.09340*, 2024. 2

[32] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 3

[33] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields, 2023. 3

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2

[35] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 3

[36] Michael Landy and J. Anthony Movshon. *The Plenoptic Function and the Elements of Early Vision*, pages 3–20. 1991. 2

[37] Donghwan Lee, Soohyun Ryu, Suyong Yeon, Yonghan Lee, Deokhwa Kim, Cheolho Han, Yohann Cabon, Philippe Weinzaepfel, Nicolas Guérin, Gabriela Csurka, and Martin Humenberger. Large-scale localization datasets in crowded indoor spaces, 2021. 1

[38] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r, 2024. 4, 8

[39] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 2, 4, 5, 6, 8, 1

[40] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation, 2022. 3

[41] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[42] Amy Lin, Jason Y. Zhang, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *CoRR*, abs/2305.04926, 2023. 6

[43] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *ICCV*, 2021. 6

[44] Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *European Conference on Computer Vision*, pages 275–292. Springer, 2022. 7

[45] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023. 7

[46] James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102 3:419–457, 1995. 3

[47] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. pages 109–165. Academic Press, 1989. 3

[48] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[49] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning, 2020. 3

[50] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5, 6, 1

[51] Jonas Ngnawe, Marianne Abemgnigni Njifon, Jonathan Heek, and Yann Dauphin. Robustmix: Improving robustness by regularizing the frequency bias of deep nets, 2024. 6

[52] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 1, 2, 4, 8, 3

[53] Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard Turner, and Mohammad Emtiyaz E Khan. Continual deep learning by functional regularisation of memorable past. In *Advances in Neural Information*

Processing Systems, pages 4453–4464. Curran Associates, Inc., 2020. 3

[54] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113: 54–71, 2019. 3

[55] Renjing Pei, Jian zhuo Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18983–18992, 2023. 2

[56] Songyou Peng, Kyle Genova, Chiyu ”Max” Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 4, 8, 3

[58] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. ICCV, 2023. 7, 2

[59] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 6

[60] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model – reduce all domains into one, 2024. 2

[61] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning, 2017. 3

[62] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 5, 6, 1, 2

[63] Pengzhen Ren, Changlin Li, Hang Xu, Yi Zhu, Guangrun Wang, Jianzhuang Liu, Xiaojun Chang, and Xiaodan Liang. Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency. *arXiv preprint arXiv:2302.10307*, 2023. 7

[64] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2022. 2

[65] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 6

[66] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research, 2019. 1

[67] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 2

[68] Jonathan Schwarz, Jelena Luketina, Wojciech M. Czarnecki, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress compress: A scalable framework for continual learning, 2018. 3

[69] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation, 2023. 3

[70] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9043–9052, 2023. 3

[71] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Kick back relax: Learning to reconstruct the world by watching slowtv, 2023. 6

[72] Julian Straub, Daniel DeTone, Tianwei Shen, Nan Yang, Chris Sweeney, and Richard Newcombe. Efm3d: A benchmark for measuring progress towards 3d egocentric foundation models, 2024. 3

[73] Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 6

[74] Ximeng Sun, Pengchuan Zhang, Peizhao Zhang, Hardik Shah, Kate Saenko, and Xide Xia. Dime-fm: Distilling multimodal and efficient foundation models, 2023. 2

[75] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5

[76] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations, 2022. 3

[77] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin

Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding, 2024. 2, 6, 7

[78] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: Solving pose estimation via diffusion-aided bundle adjustment. 2023. 6

[79] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy, 2023. 2, 4, 6

[80] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3124–3134, 2023. 2

[81] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut, 2023. 2

[82] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. CroCo v2: Improved Cross-view Completion Pretraining for Stereo Matching and Optical Flow. In *ICCV*, 2023. 2

[83] Weinzaepfel, Philippe and Leroy, Vincent and Lucas, Thomas and Brégier, Romain and Cabon, Yohann and Arora, Vaibhav and Antsfeld, Leonid and Chidlovskii, Boris and Csurka, Gabriela and Revaud Jérôme. CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion. In *NeurIPS*, 2022. 2

[84] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 7

[85] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2935–2944, 2023. 7

[86] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models, 2023. 3

[87] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation, 2024. 2

[88] Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F. Fouhey, and Joyce Chai. 3d-grand: A million-scale dataset for 3d-llms with better grounding and less hallucination, 2024. 2

[89] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding, 2021. 3

[90] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *Computer Vision and Pattern Recognition (CVPR)*, 2020. 5, 1

[91] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes, 2023. 5, 1

[92] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation, 2022. 6

[93] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning, 2024. 3

[94] Amir R. Zamir, Tilman Wekel, Pulkit Agrawal, Colin Wei, Jitendra Malik, and Silvio Savarese. *Generic 3D Representation via Pose Estimation and Matching*, page 535–553. Springer International Publishing, 2016. 1

[95] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence, 2017. 3

[96] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *ECCV*, 2022. 6

[97] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *International Conference on Learning Representations (ICLR)*, 2024. 6

[98] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *International Conference on 3D Vision*, 2022. 6

[99] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 3

[100] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 6

[101] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 3

[102] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. *ECCV*, 2024. 3

SAB3R: Semantic-Augmented Backbone in 3D Reconstruction

Supplementary Material

In Sec.A, we provide additional details about the experiments conducted in this work, including a discussion of the software used in SAB3R and a detailed breakdown of each experiment. Sec.C presents supplementary visualizations of the features generated by CLIP [57] and DINOv2 [52]. Comprehensive analysis and visualizations of our novel task, *Map and Locate*, are provided in Sec.B, including both successful and failure cases from our experiments. Finally, we discuss the limitations of our approach in Sec.D.

A. More Experiment Details

A.1. Teacher Models and Frameworks

CLIP & MaskCLIP Vision and language models are trained to generate aligned feature embeddings using a contrastive objective. The original CLIP family of models was proposed by Radford et al. [57] and included a wide variety of architectures in a private dataset of 400M image-text pairs called WIT. More recently, Ilharco et al. [29] trained several CLIP models using several architectures trained on publicly available datasets. In SAB3R , we used MaskCLIP [13], which enhances CLIP pretraining by introducing masked self-distillation. This transfers knowledge from full-image representations to masked-image predictions. This approach complements the vision-language contrastive objective by focusing on local patch representations while aligning features with indirect supervision from language. Additionally, MaskCLIP incorporates local semantic supervision into the text branch, further improving pretraining performance. We follow suggestions from FeatUp [22] that MaskCLIP [13] has better local semantic feature compare with CLIP [57].

MASt3R MASt3R [39] was trained on an extensive multi-view dataset comprising 5.3 million real-world image pairs and 1.8 million synthetic pairs. The real-world data includes diverse scenarios from ARKitScenes [1], MegaDepth [41], 3DStreetView [94], and IndoorVL [37]. The synthetic data was generated using the Habitat simulator [66], covering indoor, outdoor, and landmark environments.

Our model is finetuned on top of MASt3R, leveraging Habitat-Sim [66], ScanNet++[91], and Co3Dv2[62], ARKitScenes [1] and BlenderMVS [90].

FeatUp FeatUp [21] is a framework designed to enhance spatial resolution in deep features for tasks like segmentation and depth prediction. It addresses the loss of spatial detail caused by pooling in traditional networks using two approaches: guided upsampling with high-resolution signals in a single pass and reconstructing features at arbitrary resolutions with an implicit model. Both methods use a multi-view consistency loss inspired by NeRFs to maintain feature semantics.

FeatUp integrates seamlessly into existing pipelines, boosting resolution and performance without re-training. Experiments demonstrate its superiority over other methods in tasks such as segmentation, depth prediction, and class activation map generation. In SAB3R , we find the MaskCLIP variant of FeatUp model can also perform zero-shot semantic segmentation and we use it as our teacher model for distillation.

Table 6. **Checkpoint Details.** Information about the pre-trained checkpoints used in this work, including source and license.

Checkpoint	Source Link	License
FeatUp MaskCLIP	MaskCLIP	MIT
MASt3R	MASt3R	CC BY-NC-SA 4.0

We list the checkpoints used in SAB3R in Tab. 6, detailing the FeatUp MaskCLIP variant and MASt3R, along with their source links and license information.

A.2. Experiments Details

Monocular Depth In the main text, we benchmark SAB3R on the outdoor dataset KITTI [24] and the indoor dataset NYUv2 [50]. Here, we provide a detailed discussion of the evaluation metrics. Following DUSt3R, we use two commonly adopted metrics in monocular depth estimation:

- Absolute Relative Error (AbsRel): This measures the relative error between the ground truth depth y and the predicted depth $\hat{y}$, defined as:

$$\text{AbsRel} = \frac{|y - \hat{y}|}{y}.$$

- Prediction Threshold ($\delta_{1.25}$): This evaluates the fraction of predictions within a given threshold and is defined as:

$$\delta_{1.25} = \frac{\max\left(\frac{\hat{y}}{y}, \frac{y}{\hat{y}}\right) < 1.25}{\text{Total Predictions}}.$$

These metrics allow for comprehensive evaluation of depth prediction accuracy and robustness across different datasets.

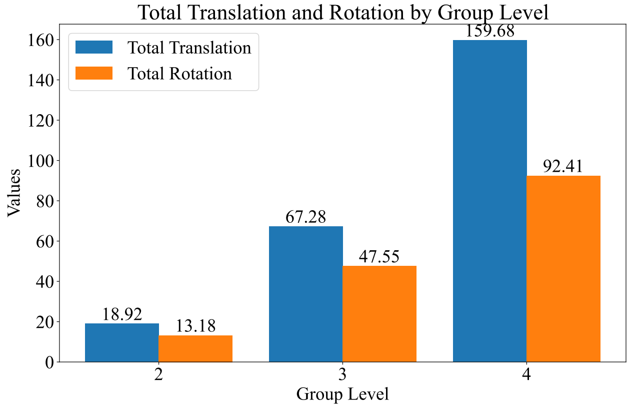

Figure 5. **Camera Distributions.** Camera translation differences and rotation differences at different group levels.

Relative Camera Pose We evaluate SAB3R on the task of relative pose estimation using the CO3Dv2 dataset [62]. To assess the relative pose error for each image pair, we report the Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA). For evaluation, we select a threshold $\tau = 15^\circ$ and report RRA@15 and RTA@15, representing the percentage of image pairs where the errors in rotation and translation are below the threshold τ.

The rotation error e_{rot} and translation error $\mathrm{e}_{\text{trans}}$ for each image pair are computed as:

$$\mathrm{e}_{\text{rot}} = \arccos\left(\frac{\text{trace}(\mathbf{R}^\top \hat{\mathbf{R}}) - 1}{2}\right),$$

$$\mathrm{e}_{\text{trans}} = \arccos\left(\frac{\mathbf{t}^\top \hat{\mathbf{t}}}{\|\mathbf{t}\|\|\hat{\mathbf{t}}\|}\right),$$

where $\mathbf{R}$ and $\hat{\mathbf{R}}$ are the ground truth and predicted rotation matrices, and $\mathbf{t}$ and $\hat{\mathbf{t}}$ are the ground truth and predicted translation vectors.

We also report the mean Average Accuracy (mAA@30), defined as the area under the accuracy curve of the angular differences for $\min(\text{RRA@30}, \text{RTA@30})$. The mAA@30 is calculated as:

$$\text{mAA@30} = \frac{1}{30}\int_0^{30} \min(\text{RRA@}\theta, \text{RTA@}\theta)\, d\theta,$$

where θ represents the threshold angle in degrees.

Zero-Shot Semantic Segmentation For zero-shot semantic segmentation, we largely follow the approach outlined by Ranasinghe et al.[58], utilizing 80 prompt templates introduced by Radford et al .[57, 77]. Class names are embedded into these prompts, and text embeddings are generated using the text encoder. We then compute the cosine similarity between each text embedding and the corresponding pixel feature—extracted directly from the CLIP head. The class with the highest cosine similarity is assigned as the predicted class for each pixel.

The class predictions are subsequently resized to match the original image dimensions, and the mean Intersection over Union (mIoU) is computed for evaluation. Unlike prior methods, our approach eliminates the concept of patches. Instead, because the CLIP head directly generates per-pixel features, we can seamlessly perform top-1 matching between semantic classes and pixel features, bypassing the need for patch-based processing.

B. Additional *Map and Locate* Details

B.1. Dataset Summary

We evaluate our *Map and Locate* framework using the ScanNet dataset [11], a large-scale indoor scene dataset that provides RGB-D sequences, camera poses, semantic and instance annotations. Specifically, we select 10 scenes from the validation split, each containing diverse object layouts and camera trajectories. Across these 10 scenes, there are a total of 436 objects with semantic and instance-level ground truth annotations.

For evaluation, we construct 2 sets of image groups for each scene, where each group comprises 2, 3, or 4 images. The image selection ensures:

- Object visibility: Objects in each group are visible across multiple images to ensure reliable localization and mapping.
- Viewpoint diversity: Selected images capture varying camera viewpoints to test robustness to occlusion and perspective changes.

In total, this results in 60 image groups (2 sets per scene × 10 scenes × 3 group sizes). Each group is paired with its corresponding rgb images, depth maps, camera poses (intrinsics and extrinsic) , and semantic and instance labels, providing a comprehensive benchmark for evaluating both mapping accuracy and object localization performance.

B.2. Dataset Visualizations

We present a dataset statistics visualization in Fig. 5, showing camera translation differences and rotation differences. Translation differences are computed as the Euclidean distance between translation vectors, $d_{\text{translation}} = \|\mathbf{t}_1 - \mathbf{t}_2\|_2$, and rotation differences are calculated as the geodesic distance on $SO(3)$, $d_{\text{rotation}} = \|\mathbf{r}_\Delta\|_2$, where $\mathbf{r}_\Delta$ is the axis-angle representation of the relative rotation $\mathbf{R}_\Delta = \mathbf{R}_1^{-1}\mathbf{R}_2$. These metrics highlight the variability in camera poses across the dataset. We observe that as the number of views increases, both camera translation differences and rotation differences grow. Despite this, our results demonstrate consistent performance across all group levels, highlighting the robustness of our algorithm.

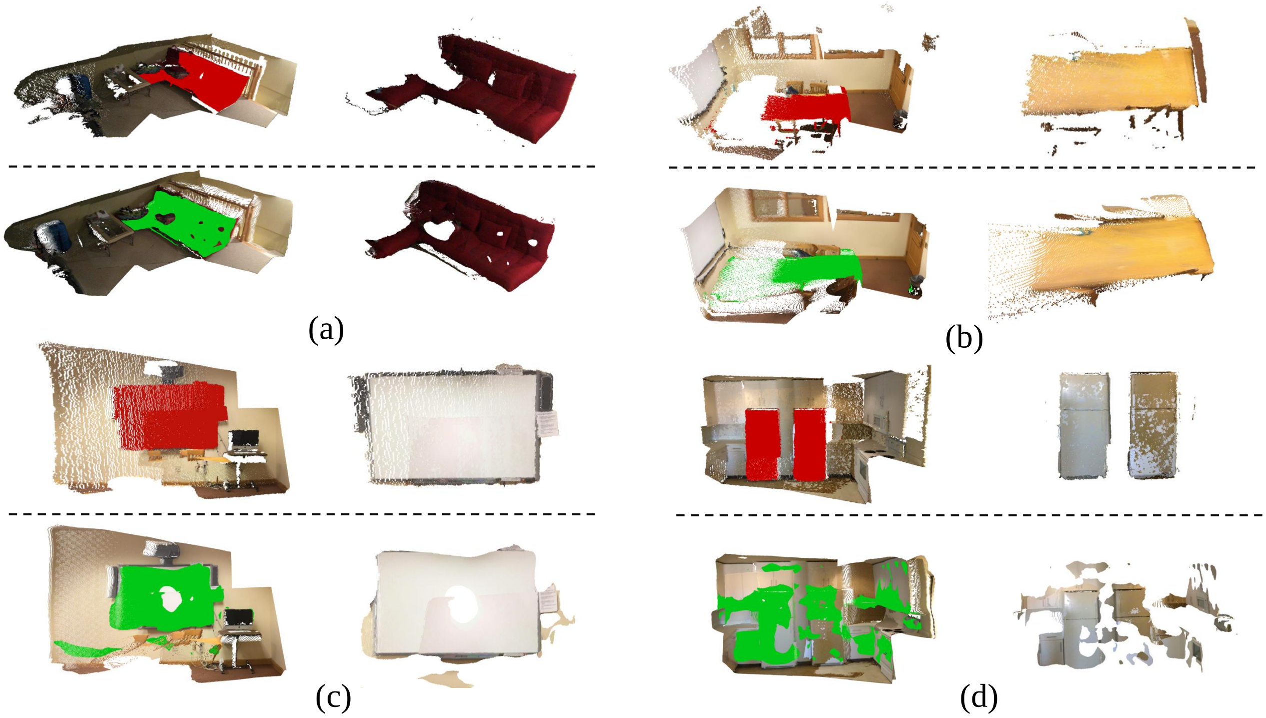

Figure 6. **Qualitative Examples of *Map and Locate* with `SAB3R`.** Panels (a), (b), and (c) illustrate successful examples of 3D scene reconstruction and accurate object segmentation. In each sub-group, the top row shows the ground truth, with the target objects highlighted in red, accompanied by visualizations of segmented objects for each ground truth target. The bottom row presents the predicted results, where the segmented objects are shown in green, with the extracted objects displayed on the right for clarity. Panel (d) provides an example of a failure case.

B.3. More Qualitative Examples

Fig. 6 presents additional qualitative examples demonstrating the performance of *Map and Locate* with `SAB3R`.

C. Additional visualization

Fig. 7 presents additional visualizations of 3D features from DINO [52] and CLIP [57]. The visualizations highlight distinct features for different objects. Predicted RGB is provided as a reference.

D. Limitations

Our study is constrained by limited computational resources, which restricted us from training the model for more epochs, potentially resulting in under-trained checkpoints. Additionally, predicting dense features significantly increases vRAM requirements, further limiting our ability to optimize the model fully. Due to these resource constraints, we were unable to use the entire pre-training dataset for fine-tuning, which may have prevented the model from achieving its best possible performance. Our novel task, *Map and Locate*, relies on the ScanNet dataset, which, despite its comprehensiveness, is primarily biased toward indoor environments. Extending this work to more diverse datasets, including outdoor or dynamic scenes, represents an interesting direction for future works.

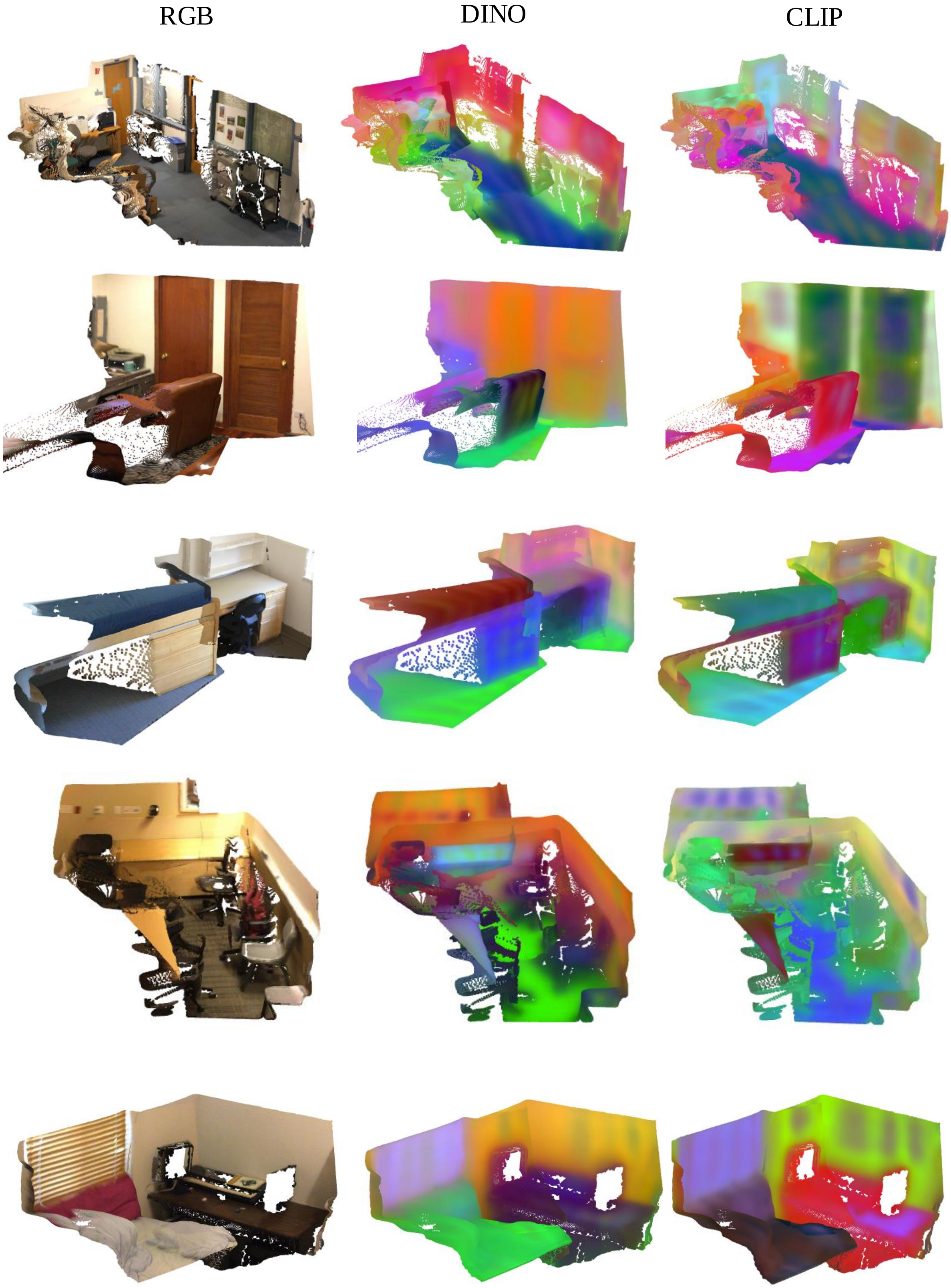

Figure 7. **3D Feature Visualizations.** Additional visualizations of 3D features are presented for DINO and CLIP, alongside the original RGB 3D point map for reference.