

PE3R: Perception-Efficient 3D Reconstruction

Jie Hu, Shizun Wang, Xinchao Wang*
xML Lab, National University of Singapore

Abstract

Recent advances in 2D-to-3D perception have enabled the recovery of 3D scene semantics from unposed images. However, prevailing methods often suffer from limited generalization, reliance on per-scene optimization, and semantic inconsistencies across viewpoints. To address these limitations, we introduce PE3R, a tuning-free framework for efficient and generalizable 3D semantic reconstruction. By integrating multi-view geometry with 2D semantic priors in a feed-forward pipeline, PE3R achieves zero-shot generalization across diverse scenes and object categories without any scene-specific fine-tuning. Extensive evaluations on open-vocabulary segmentation and multi-view depth estimation show that PE3R not only achieves up to $9\times$ faster inference but also sets new state-of-the-art accuracy in both semantic and geometric metrics. Our approach paves the way for scalable, language-driven 3D scene understanding. Code is available at github.com/hujiecpp/PE3R.

1. Introduction

While machine vision has achieved remarkable success in 2D perception tasks, human perception is inherently three-dimensional, seamlessly integrating multiple viewpoints to construct a coherent understanding of the world [2, 43, 54]. This fundamental gap motivates a central challenge in computer vision: can we harness powerful 2D perception models to reconstruct and comprehend 3D scenes without relying on 3D supervision? The problem is twofold: it requires not only estimating accurate geometry from unposed images but also lifting 2D semantic information into a consistent 3D representation.

Recent methods have made significant strides in 2D-to-3D perception, recovering both geometry and semantics from unposed multi-view images [5, 6, 15, 17, 18, 20, 23, 27, 32, 33, 45–47, 56, 62]. Despite this progress, these approaches often face a trilemma: they struggle with scene generalization, fail to maintain semantic consistency across views, and suffer from high computational costs. For instance, Neural Radiance Fields (NeRF) [29] and 3D Gaus-

sian Splatting (3DGS) [19] can produce high-quality reconstructions but typically require per-scene training and involve heavy semantic lifting, making them impractical for real-time or large-scale applications.

To overcome these limitations, we introduce PE3R, a tuning-free framework for efficient, accurate, and generalizable 3D semantic reconstruction. Inspired by recent advances in feed-forward geometry estimation [24, 50, 52], PE3R employs a feed-forward pipeline that eliminates scene-specific fine-tuning while enabling rapid inference. As illustrated in Fig. 1, our framework integrates three carefully designed modules: (1) *Pixel Embedding Disambiguation* ensures semantic consistency across views and object hierarchies; (2) *Semantic Point Cloud Reconstruction* fuses geometry with dense 2D semantics to refine 3D structure; and (3) *Global View Perception* enables open-vocabulary object retrieval through language grounding. This cohesive design empowers robust zero-shot generalization to previously unseen environments.

We conduct extensive evaluation across diverse benchmarks, including Mip-NeRF360 [3], Replica [44], ScanNet++ [57] for open-vocabulary segmentation, and KITTI [14], ScanNet [9], DTU [1], ETH3D [39], Tanks and Temples [22] for multi-view depth estimation. Results demonstrate that PE3R achieves up to $9\times$ faster 3D semantic reconstruction than prior methods while advancing in both semantic fidelity and geometric precision.

Our main contributions are summarized as follows:

- We propose PE3R, a fast and scalable framework for 3D semantic reconstruction that operates directly on unposed 2D images, requiring no 3D supervision, per-scene fine-tuning, or camera parameters.
- We introduce three novel modules that address key challenges in 2D-to-3D perception: pixel embedding disambiguation for cross-view consistency, semantic point cloud reconstruction for geometry-semantics fusion, and global view perception for open-vocabulary interaction.
- Through comprehensive evaluations on multiple benchmarks, we demonstrate state-of-the-art performance, effective zero-shot generalization, and practical scalability. Code is provided to ensure reproducibility.

*Corresponding author: xinchao@nus.edu.sg



Figure 1. **Examples of Perception-Efficient 3D Reconstruction.** Our framework reconstructs 3D semantic scenes from unposed RGB images and supports open-vocabulary interaction, without any training or per-scene optimization. It is efficient in two critical ways: (1) *Input efficiency*: no depth maps or camera parameters are needed; (2) *Inference efficiency*: significant speed gains on semantic reconstruction compared to prior pipelines. These properties make PE3R practical for real-world deployment in large-scale and time-sensitive scenarios.

2. Related Work

2D-to-3D Reconstruction. Significant progress has been made in reconstructing 3D scenes from 2D images, with many methods achieving high surface fidelity and geometric coherence. A number of approaches leverage signed distance functions (SDFs) combined with volume rendering for accurate mesh extraction [13, 16, 28, 51, 53]. Neural Radiance Fields (NeRF) [29] popularized the use of implicit neural representations for view synthesis, while 3D Gaussian Splatting (3DGS) [19] later introduced an explicit, point-based alternative that enables highly efficient rendering. To overcome the dependency on known camera parameters, several recent works operate directly on unposed images. Among them, DUST3R [52] predicts 3D structure from image sets. MAST3R [24] extends this direction by incorporating keypoint correspondences, and MAST3R-SfM [10] further unifies structure-from-motion and multi-view stereo within a transformer framework. Hybrid techniques such as Vggsfm [49] combine traditional geometry with learned features to improve robustness. Scalable systems like Spann3R [48], FAS3R [55], and VGGT [50] integrate attention mechanisms or visibility reasoning to enhance multi-view consistency and computational efficiency. While these methods achieve impressive geometric results, they typically lack integrated semantic understanding or require task-specific adaptation. PE3R builds upon these geometric foundations but moves beyond them by incorporating semantic priors, enabling efficient and context-aware 3D scene understanding without per-scene training.

2D-to-3D Perception. A parallel line of research seeks to lift 2D semantic information into 3D reconstructions. Early works such as GNeRF [7] and InPlace [61] embed-

ded 2D masks into NeRF representations to enable 3D-aware segmentation with 2D supervision. Subsequent efforts extended NeRF to support 3D instance segmentation, using either supervised [4, 26, 42] or unsupervised [30, 58] learning paradigms. The introduction of the Segment Anything Model (SAM) [21] inspired a range of SAM-guided 3D lifting pipelines, including SA3D [5] and Feature3DGS [62], which incorporate prompt-driven segmentation into 3D scene representations. Several other methods directly project SAM masks into 3D space, such as SAGA [6], Gaussian Grouping [56], SAGS [18], Click-Gaussian [8], and FlashSplat [41]. The advent of vision-language models has further enabled open-vocabulary 3D segmentation. For example, LangSplat [33] aligns CLIP embeddings with 3D Gaussians, while GOI [34] enforces multi-view consistency through text-image alignment. The Large Spatial Model (LSM) [12] performs open-vocabulary semantic segmentation by lifting features from a pre-trained 2D model (LSeg) and aggregating them into a unified 3D representation. A fundamental limitation shared by most of these methods is their reliance on scene-specific training, iterative semantic optimization, or manual heuristics, which constrains their generalization and practical efficiency. In contrast, PE3R supports scalable, zero-shot 3D semantic reconstruction through a cohesive feed-forward pipeline that jointly optimizes geometry and semantics, eliminating the need for 3D supervision or per-scene training.

2D Foundational Models. The advancement of large-scale pre-trained models has been pivotal for generalizable visual perception. CLIP [35] established a shared image-text embedding space through contrastive learning, with SigLIP [59] offering a more efficient variant. SAM [21] enables promptable segmentation of arbitrary image re-

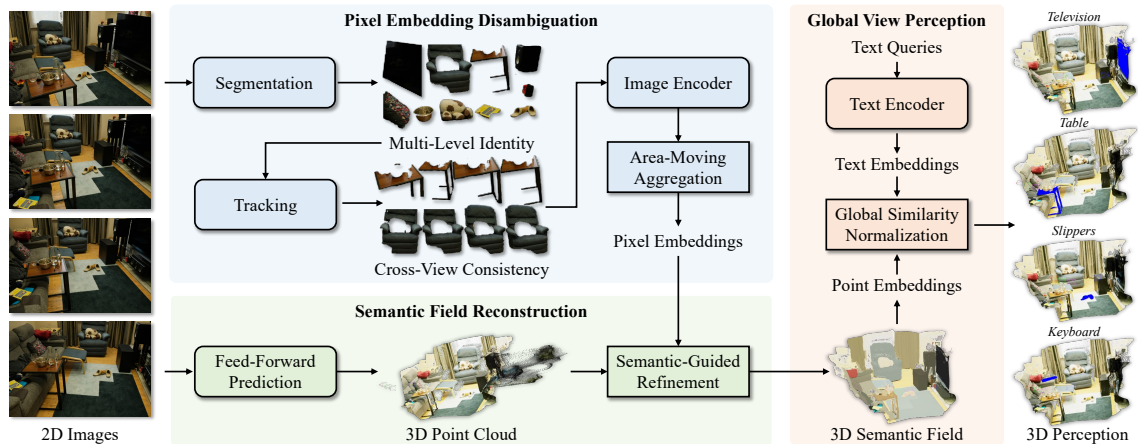


Figure 2. **The PE3R Framework.** Our pipeline comprises three core stages: (1) Pixel Embedding Disambiguation: Images are decomposed into hierarchical masks (via SAM/SAM2), which are encoded by CLIP and aggregated to yield consistent per-pixel embeddings. (2) Semantic Point Cloud Reconstruction: A feed-forward model (DUSt3R) generates 3D point clouds, which are then refined using semantic cues. (3) Global View Perception: Text embeddings are compared against the 3D point features via global similarity normalization, enabling open-vocabulary semantic localization.

gions, with SAM2 [36] enhancing cross-view and temporal consistency. DINOv2 [31] delivers high-quality visual features through self-supervised learning, and Grounding DINO [25] facilitates open-set object detection using language-based supervision. PE3R strategically leverages these foundational models not merely as feature extractors, but as core components within a novel pipeline designed to generate robust, language-grounded 3D semantic representations, all without requiring any 3D model fine-tuning.

3. Method

3.1. Problem Formulation

We address the challenging problem of reconstructing a semantically rich 3D scene representation from a collection of 2D images captured in unconstrained, in-the-wild settings. The core difficulty stems from the absence of any geometric supervision of semantics. To bridge this gap, we aim to reconstruct a 3D semantic point cloud directly from a set of unposed 2D images, capturing both the spatial structure and high-level semantic content of the scene. Our framework is designed to operate without any 3D semantic annotations, camera intrinsics, or extrinsics, ensuring scalability and practicality in unconstrained environments. The reconstructed scene supports natural language interaction, allowing users to issue open-ended text queries (e.g., “black chair”, “white bottle”) to retrieve and localize corresponding semantic entities in 3D, thereby enabling interactive and human-aligned perception.

3.2. PE3R Framework

As illustrated in Fig. 2, the PE3R pipeline is architected around three cohesive stages that collectively address the

challenge of 3D semantic reconstruction from unposed images. (1) Pixel Embedding Disambiguation: This stage focuses on deriving view-consistent and semantically coherent embeddings from the input images. Each image is first decomposed into hierarchical object masks using SAM [21] or SAM2 [36]. These masks are encoded by a vision-language model (e.g., CLIP [35]) into a shared semantic space. The core of this module lies in an area-weighted spherical interpolation strategy that aggregates region features across different views and hierarchical levels. This process resolves semantic ambiguities caused by occlusions or nested objects, yielding dense, per-pixel semantic embeddings that are consistent across images. (2) Semantic Point Cloud Reconstruction: Building upon the disambiguated 2D embeddings, this stage transitions to 3D geometric reconstruction. A feed-forward model, such as DUSt3R [52], is leveraged to generate an initial dense 3D point cloud from the multi-view images. The key innovation here is the semantic-guided refinement of this raw geometry. The point cloud is refined by leveraging the previously computed semantic embeddings to filter spatial outliers and enhance coherence, resulting in a unified reconstruction that is both geometrically accurate and semantically labeled. (3) Global View Perception: The final stage enables open-vocabulary interaction with the reconstructed 3D scene. A user’s natural language query is first encoded into a textual embedding using CLIP’s text encoder. This query embedding is then compared against the semantic features of every 3D point via cosine similarity. To ensure a consistent and interpretable retrieval process across all views, a global min-max normalization is applied to the similarity scores. 3D points whose normalized similarity scores exceed a predefined threshold are retrieved and

precisely localized, thereby facilitating real-time, language-grounded understanding and exploration of the scene.

3.3. Pixel Embedding Disambiguation

Establishing a consistent semantic embedding space across multiple views is foundational for robust 3D reconstruction. This process, however, is inherently challenged by two types of ambiguity: (i) *object-level ambiguity*, where overlapping or nested structures (*e.g.*, a cup on a table) create uncertain semantic associations; and (ii) *viewpoint-level ambiguity*, where occlusions or perspective changes lead to inconsistent semantics across different viewpoints. To resolve these issues, we construct dense per-pixel embeddings that are both discriminative at the object level and consistent across all views. In our formulation, an “object” is defined as a collection of hierarchically related masks with substantial spatial overlap. Specifically, a smaller mask is grouped into a larger object if their Intersection over Union (IoU) exceeds a threshold of 0.9. This simple yet effective heuristic allows us to associate fine-grained parts (*e.g.*, chair legs) with their corresponding whole-object regions, without requiring any external supervision or predefined labels.

Image Embedding Extraction. Given n input images $\mathbf{X}^1, \dots, \mathbf{X}^n \in \mathbb{R}^{3 \times H \times W}$, we employ SAM [21] and SAM2 [36] to decompose each image into a hierarchy of object masks with consistent object indices across views. This yields a collection of masked image regions $\mathbf{M}^1, \dots, \mathbf{M}^n \in \mathbb{R}^{m \times 3 \times H \times W}$, where m is the number of masks per image. A vision-language encoder $\mathcal{F}_{img}(\cdot)$ is used to extract each masked region into the semantic space:

$$\mathbf{F}^1, \mathbf{F}^2, \dots, \mathbf{F}^n = \mathcal{N}(\mathcal{F}_{img}(\mathbf{M}^1, \mathbf{M}^2, \dots, \mathbf{M}^n)), \quad (1)$$

where $\mathbf{F}^i \in \mathbb{R}^{m \times d}$ denotes the L2-normalized ($\mathcal{N}(\cdot)$) embedding matrix for the i -th image.

Area-Weighted Interpolation. To resolve semantic ambiguities and align features across both views and hierarchy levels, we introduce a geometry-aware aggregation strategy. Our intuition is that masks with larger visible areas typically provide more stable and reliable semantic features, and thus should exert a greater influence on the aggregated representation. Given two unit embeddings \mathbf{F}_A and \mathbf{F}_B with corresponding areas area_A and area_B , we define the area ratio $t = \frac{\text{area}_B}{\text{area}_A + \text{area}_B}$ and compute the aggregated embedding as:

$$\hat{\mathbf{F}}_B = a\mathbf{F}_A + b\mathbf{F}_B, a = \frac{\sin((1-t)\theta)}{\sin(\theta)}, b = \frac{\sin(t\theta)}{\sin(\theta)}, \quad (2)$$

where θ is the angle between \mathbf{F}_A and \mathbf{F}_B . This formulation possesses two key properties that underpin its effectiveness:

Proposition 1. Vector Normalization: The interpolated vector $\hat{\mathbf{F}}_B$ preserves unit norm, ensuring it remains within the original semantic embedding space.

Dataset	Method	mIoU	mPA	mP
Mip.	LERF [20]	0.2698	0.8183	0.6553
	F-3DGS [62]	0.3889	0.8279	0.7085
	GS Grouping [56]	0.4410	0.7586	0.7611
	LangSplat [33]	0.5545	0.8071	0.8600
	OpenNeRF [11]	0.5734	0.8345	0.8959
	LSM [12]	0.6254	0.8887	0.9192
	GOI [34]	0.8646	0.9569	0.9362
	PE3R, ours	0.8951	0.9617	0.9726
Rep.	LERF [20]	0.2815	0.7071	0.6602
	F-3DGS [62]	0.4480	0.7901	0.7310
	GS Grouping [56]	0.4170	0.7370	0.7276
	LangSplat [33]	0.4703	0.7694	0.7604
	OpenNeRF [11]	0.5014	0.7915	0.7831
	LSM [12]	0.5614	0.8113	0.7961
	GOI [34]	0.6169	0.8367	0.8088
	PE3R, ours	0.6531	0.8377	0.8444

Table 1. **2D-to-3D Open-Vocabulary Segmentation** results on small-scale Mip-NeRF360 (Mip.) and Replica (Rep.) datasets.

Proposition 2. Semantic Guidance: If \mathbf{F}_A has a higher similarity to a reference semantic vector \mathbf{F}_C than \mathbf{F}_B does, then $\hat{\mathbf{F}}_B$ will also exhibit a higher similarity to \mathbf{F}_C than \mathbf{F}_B does. This steers the aggregated representation toward more semantically meaningful directions.

These properties guarantee that our interpolation strategy is both geometrically sound and semantically informative.

Pixel Embedding Ensemble. To enhance semantic coherence, we perform a two-stage ensemble that operates both within and across views. (1) Within-view aggregation: For each image, object masks are processed in descending order of area. The embeddings of smaller masks (likely object parts) are iteratively aggregated via area-weighted interpolation with those of larger masks (likely whole objects), promoting semantic consistency across the hierarchical part-whole structure within a single view. (2) Cross-view aggregation: We seed the SAM2 tracker on the first view using SAM (segment-everything) masks; for each subsequent view, SAM2 propagates tracked masks and SAM proposes all masks again. Any SAM mask with $\text{IoU} < 0.1$ to all current SAM2 masks is treated as newly discovered/re-detected and inserted as a new tracking prompt (improving recall under tracking loss). When association is unreliable, we skip cross-view fusion and fall back to within-view disambiguation. This two-level ensemble produces a set of stable and semantically coherent object embeddings that are consistent across both spatial hierarchies and multiple views. The final object-level embeddings are then projected back into the pixel space, yielding dense per-pixel semantic maps $\mathbf{E}^1, \mathbf{E}^2, \dots, \mathbf{E}^n \in \mathbb{R}^{H \times W \times d}$.

3.4. Semantic Point Cloud Reconstruction

Building upon the disambiguated pixel embeddings, this module focuses on reconstructing a high-quality 3D semantic point cloud. We leverage recent feed-forward 3D pointmap predictors, such as DUST3R [52], to directly estimate per-pixel 3D coordinates (or pointmaps) from the

Method	Preprocess	Training	Total
LERF [20]	3mins	40mins	43mins
F-3DGS [62]	25mins	623mins	648mins
GS Grouping [56]	27mins	138mins	165mins
LangSplat [33]	50mins	99mins	149mins
GOI [34]	8mins	37mins	45mins
PE3R, ours	5mins	-	5mins

Table 2. **Runtime Comparison for 3D Semantic Reconstruction** on Mipnerf360.

Method	mIoU	mPA	mP
LERF [20] Embedding	0.1824	0.6024	0.5873
GOI [34] Embedding	0.2101	0.6216	0.6013
LSM [12] Embedding	0.2124	0.6346	0.6114
PE3R Embedding (ours)	0.2248	0.6542	0.6315

Table 3. **2D-to-3D Open-Vocabulary Segmentation** results on the large-scale ScanNet++ dataset.

multi-view images:

$$\mathbf{P}^1, \mathbf{P}^2, \dots, \mathbf{P}^n = \mathcal{F}_{pts}(\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^n), \quad (3)$$

where each $\mathbf{P}^i \in \mathbb{R}^{H \times W \times 3}$ represents the 3D coordinates (x, y, z) for every pixel in the i -th view. While this feed-forward paradigm is fast, the initial pointmaps are often contaminated by spatial noise due to visual artifacts like occlusions, reflections, and transparent surfaces. To overcome this inherent limitation, we introduce a semantic-guided refinement process that utilizes the consistent embeddings to enforce both geometric and semantic coherence.

Anomaly Point Detection. The core premise of our detection strategy is that pixels belonging to the semantic region should form a spatially coherent cluster in 3D. We thus identify outliers by evaluating the local 3D consistency within semantically homogeneous areas. For each pixel $P_{i,j}$, we compute its average Euclidean distance to neighboring pixels within a $k \times k$ window that share the same semantic label:

$$L_{i,j} = \frac{\sum_{dx,dy} \mathcal{I}(\mathbf{M}_{i,j}, \mathbf{M}_{i+dx,j+dy}) \cdot \mathcal{D}(P_{i,j}, P_{i+dx,j+dy})}{\sum_{dx,dy} \mathcal{I}(\mathbf{M}_{i,j}, \mathbf{M}_{i+dx,j+dy})}, \quad (4)$$

where $dx, dy \in [-\lfloor k/2 \rfloor, \lfloor k/2 \rfloor]$, $\mathcal{I}(\cdot, \cdot)$ is an indicator function that evaluates to 1 if two pixels share the same instance-track index, and $\mathcal{D}(\cdot, \cdot)$ denotes the Euclidean distance in 3D space. This operation yields a semantic-aware distance map $\mathbf{L}^i \in \mathbb{R}^{H \times W}$ for each view. Pixels associated with distance values exceeding a predefined threshold after normalization are flagged as spatial anomalies and subsequently excluded, resulting in a significantly cleaner intermediate point cloud.

Semantic-Guided Refinement. Merely filtering outliers, however, does not actively correct the underlying erroneous estimates. Instead of applying computationally expensive post-hoc geometric regularization (e.g., least-squares fit-

ting), we propose an efficient image-space smoothing strategy that tackles the problem at its root. For each detected anomaly, we adjust its RGB value \mathbf{y} by blending it with the average color \mathbf{x} of its surrounding semantic region:

$$\hat{y} = \alpha \cdot x + (1 - \alpha) \cdot y. \quad (5)$$

Here, the blending factor $\alpha \in [0, 1]$ controls the strength of semantic smoothing, effectively suppressing textural noise that misleads the geometric predictor while preserving genuine structural edges. The smoothed image is then fed back into the pointmap predictor \mathcal{F}_{pts} , yielding a refined 3D estimate that is more spatially coherent and better aligned with the semantic structure of the scene. Finally, a global alignment step synchronizes all refined pointmaps $\hat{\mathbf{P}}^i$ into a unified coordinate frame, ensuring multi-view consistency. The output is a fused semantic point cloud where each point is associated with a 3D coordinate and its corresponding semantic embedding, forming a robust foundation for language-grounded interaction.

3.5. Global View Perception

The final stage of PE3R leverages the unified semantic point cloud to enable intuitive, open-vocabulary interaction with the reconstructed 3D semantic scene. Given a user-provided text query, we first encode it into a semantic embedding using a language encoder $\mathcal{F}_{txt}(\cdot)$ (e.g., CLIP [35]):

$$\mathbf{T} = \mathcal{F}_{txt}(\text{text}), \quad (6)$$

where $\mathbf{T} \in \mathbb{R}^d$ resides in the shared image-language embedding space. To identify corresponding regions in 3D, we compute the cosine similarity between the query embedding \mathbf{T} and the dense per-pixel semantic embeddings $\mathbf{E}^1, \mathbf{E}^2, \dots, \mathbf{E}^n$ from all input views:

$$[\mathbf{S}^1, \mathbf{S}^2, \dots, \mathbf{S}^n] = \mathcal{D}(\mathbf{T}, [\mathbf{E}^1, \mathbf{E}^2, \dots, \mathbf{E}^n]), \quad (7)$$

where $\mathcal{D}(\cdot)$ denotes the cosine similarity function. To ensure consistency in retrieval scores across different viewpoints, we apply a global min-max normalization to all similarity values. Specifically, we concatenate the similarity scores from every view and scale them globally into the range $[0, 1]$. This provides a unified and interpretable similarity measure that is invariant to view-specific variations. A predefined threshold is then applied to the normalized scores to select the most relevant 3D points from the corresponding refined pointmaps $\mathbf{P}^1, \mathbf{P}^2, \dots, \mathbf{P}^n$. Points whose normalized similarity scores exceed this threshold are considered semantically aligned with the text query and are retained for 3D localization. This approach enables natural language-based interaction with the reconstructed 3D scene, supporting flexible object retrieval and semantic scene understanding without the need for predefined category labels or model retraining.

Methods	KITTI		ScanNet		ETH3D		DTU		T&T		Ave.	
	rel↓	τ ↑	rel↓	τ ↑	rel↓	τ ↑	rel↓	τ ↑	rel↓	τ ↑	rel↓	τ ↑
(a) COLMAP [37]	12.0	58.2	14.6	34.2	16.4	55.1	0.7	96.5	2.7	95.0	9.3	67.8
	26.9	52.7	38.0	22.5	89.8	23.2	20.8	69.3	25.7	76.4	40.2	48.8
(b) DUS _t 3R [52]	9.1	39.5	4.9	60.2	2.9	76.9	3.5	69.3	3.2	76.7	4.7	64.5
	-	-	6.3	50.3	4.7	62.7	3.9	62.6	4.4	64.0	-	-
	11.0	33.2	4.8	60.3	3.1	74.5	2.7	75.7	2.9	78.5	4.9	64.4
	36.9	5.4	22.0	9.6	27.9	9.9	13.6	13.7	22.1	14.6	24.5	10.6
	9.4	41.3	4.4	65.0	1.8	86.3	1.1	94.3	2.1	85.2	3.8	74.4
(c) PE3R, with DUS _t 3R	9.4	48.6	5.5	55.1	2.3	82.0	3.2	69.1	2.1	85.3	4.5	68.0
	9.2	43.8	4.3	65.8	1.7	86.6	1.1	94.4	2.0	85.4	3.7	75.2

Table 4. **Multi-view Depth Estimation on RobustMVD.** Comparison under: (a) classical methods, (b) feed-forward methods, and (c) semantic reconstruction. PE3R achieves the best average performance, demonstrating that semantic guidance generally enhances geometric accuracy. † indicates our implementation. Metrics are indicated by relative error (rel ↓) and inlier ratio (τ ↑).

4. Experiments

4.1. Experimental Details

Datasets and Tasks. We evaluate PE3R on two core tasks to comprehensively assess its capabilities: *open-vocabulary 2D-to-3D segmentation* and *multi-view depth estimation*. For open-vocabulary segmentation, we employ three benchmarks: Mip-NeRF360 [3] and Replica [44] are used to evaluate performance in relatively compact, object-centric scenes, using the open-vocabulary labels from GOI [34]. ScanNet++ [57] is adopted to test generalization in large-scale, complex indoor environments with diverse layouts and object categories. For multi-view depth estimation, we follow established benchmarks to assess geometric accuracy across a wide range of scenarios. The evaluated datasets include KITTI [14] (outdoor driving), ScanNet [9] (indoor scenes), DTU [1] (controlled object-level scans), ETH3D [39] (high-resolution scenes), and Tanks and Temples (T&T) [22] (large-scale outdoor). This selection ensures a comprehensive evaluation across both indoor and outdoor environments.

Implementation Details. PE3R is built upon several powerful, off-the-shelf models. We employ MobileSAMv2 [60] for efficient image segmentation, SAM2 [36] for obtaining temporally consistent mask tracks, and DUS_t3R [52] as our core feed-forward 3D pointmap predictor. We use SAM for per-image “segment-everything” masks, and SAM2 for instance tracking. Concretely, we initialize SAM2 on the first image by running SAM and using the resulting masks as prompts to seed the SAM2 tracker. For each subsequent image, we (i) run SAM2 to propagate tracked masks, then (ii) run SAM to re-detect all objects; any SAM mask with $\text{IoU} < 0.1$ to all current SAM2 masks is treated as newly discovered/re-detected and added as a new instance to the tracker, recovering missed/lost tracks under large viewpoint changes. All experiments were conducted on a single server equipped with an NVIDIA A100 GPU and an Intel i7 CPU.

Evaluation Protocol. Our evaluation protocol is designed to be fair, reproducible, and aligned with standard prac-

tices in the field. For open-vocabulary segmentation, we report the standard metrics of mean Intersection over Union (mIoU), mean Pixel Accuracy (mPA), and mean Precision (mP). The evaluation procedure is tailored to dataset characteristics to ensure optimal and fair benchmarking. On Mip-NeRF360 and Replica, we follow the protocol established by GOI [34]. We perform a holistic 3D semantic reconstruction of the entire scene. Open-vocabulary queries, labeled by GOI, are then executed in this unified 3D space. The resulting 3D segmentation is projected back onto the original 2D image planes for quantitative comparison against the 2D ground-truth masks. On the larger ScanNet++ dataset, we process data in snippets of four consecutive frames to reduce the computational load, a common practice for scaling 3D methods to large scenes. This approach effectively highlights the benefit of integrating semantic features into 3D reconstruction. For a direct and fair comparison, we integrate the semantic embeddings from baseline methods, LERF [20], LSM [12], and GOI [34], directly into our pipeline. This is done by replacing *only* the pixel-level semantic embeddings while keeping the geometry backbone and the entire evaluation procedure identical. The final evaluation is conducted by projecting the 3D segmentation results back to 2D for comparison with ground-truth annotations. This controlled setup ensures that performance differences are attributable solely to the quality of the semantic embeddings. We report the total pipeline processing time (preprocessing and inference) to assess computational efficiency. For multi-view depth estimation, we adhere to the official protocol of the RobustMVD benchmark [40]. For each test sample, we use source views selected via the quasi-optimal strategy. Performance is measured using Absolute Relative Error (Rel) ↓ and the Inlier Ratio (τ) ↑ at a threshold of 1.03. Results are reported for each individual dataset as well as the overall average across all benchmarks.

4.2. Main Results

2D-to-3D Open-Vocabulary Segmentation. We first evaluate PE3R on the Mip-NeRF360 and Replica datasets,

Method	mIoU	mPA	mP
w/o Multi-Level Disam.	0.1624	0.5892	0.5623
w/o Cross-View Disam.	0.1895	0.6012	0.5923
w/o Global MinMax Norm.	0.2035	0.6253	0.6186
PE3R (full)	0.2248	0.6542	0.6315

Table 5. **Effect of Disambiguation Modules** for 2D-to-3D open-vocabulary segmentation on ScanNet++.

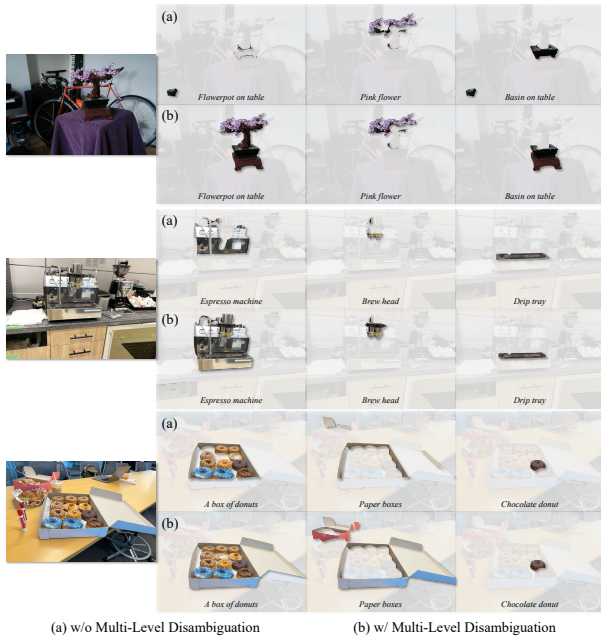


Figure 3. **Effect of Multi-Level Disambiguation.** Without this module, the model fails to associate object parts with the whole, leading to fragmented semantics.

which provide a standard benchmark for comparing with state-of-the-art NeRF- and 3DGS-based methods. As summarized in Table 1, PE3R consistently surpasses all previous approaches, establishing a new state-of-the-art across all metrics (mIoU, mPA, mP). A pivotal advantage of our framework is its exceptional computational efficiency. As evidenced by Table 2, PE3R completes the full semantic reconstruction pipeline in merely 5 minutes. This represents up to a $9\times$ speedup over optimization-based methods. To assess generalization, we evaluate PE3R on the large-scale and complex ScanNet++ benchmark. Given the scalability constraints of methods requiring per-scene optimization, we conduct a direct and fair comparison by integrating the 2D semantic features from LERF, LSM, and GOI into our pipeline under identical settings. As reported in Table 3, PE3R achieves superior segmentation accuracy, confirming its ability to maintain high performance while scaling effectively to challenging, large-scale environments. Qualitative results in Fig. 1 further corroborate these findings, showcasing robust and precise segmentation across diverse indoor, outdoor, and in-the-wild scenes.

Multi-view Depth Estimation. We further evaluate the ge-

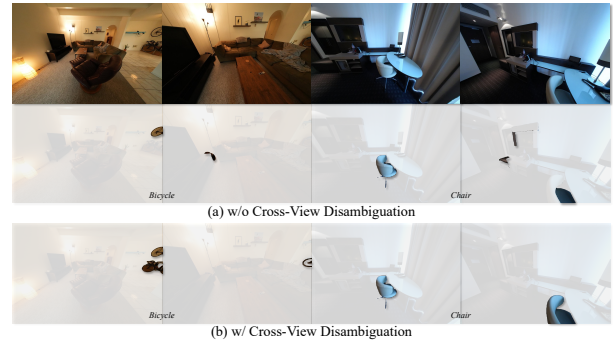


Figure 4. **Effect of Cross-View Disambiguation.** In the absence of cross-view alignment, the same object is inconsistently labeled across viewpoints.

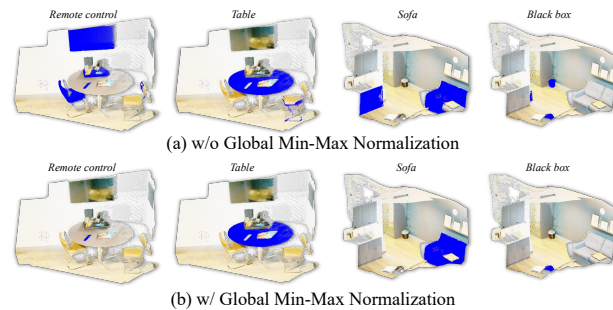


Figure 5. **Effect of Global Min-Max Normalization.** Without global normalization, similarity scores become uncalibrated across views, impairing retrieval reliability.

ometric fidelity of PE3R on the task of multi-view depth estimation. For a comprehensive analysis, Table 4 organizes methods into three categories: (a) classical 3D reconstruction approaches, (b) modern feed-forward methods, and (c) our full PE3R framework with semantic-guided reconstruction. PE3R demonstrates highly competitive performance across a wide spectrum of datasets, achieving the best overall average on standard metrics (Rel and τ). This indicates that the integration of semantic priors generally enhances the robustness of the geometric reconstruction. An in-depth analysis reveals that the benefits of semantic guidance are scene-dependent. While PE3R shows notable improvements on most benchmarks, its performance on ScanNet and DTU is marginally behind the DUST3R baseline. We attribute this to the characteristic challenges of these datasets, such as low semantic diversity, repetitive textures, and limited viewpoint coverage, where semantic cues can introduce ambiguity rather than resolve it. This nuanced performance profile highlights a key insight: semantic guidance is most beneficial in semantically rich and visually distinct environments. Nevertheless, the improvement in the overall benchmark average strongly validates that our semantic augmentation strategy provides a net positive effect on reconstruction quality.

Multi-Level Disam.	Cross-View Disam.	mIoU	mPA	mP
averaged	averaged	0.1994	0.6001	0.5985
averaged	area-slerp	0.2158	0.6141	0.6015
area-slerp	averaged	0.2058	0.6041	0.6003
conf-weighted	conf-weighted	0.1532	0.5723	0.5435
conf-weighted	conf-slerp	0.1743	0.5842	0.5554
conf-slerp	conf-weighted	0.1721	0.5813	0.5493
conf-slerp	conf-slerp	0.1998	0.5991	0.5918
area-slerp	area-slerp	0.2248	0.6542	0.6315

Table 6. Comparison of Feature Aggregation Strategies on ScanNet++.

4.3. Ablation Studies

To provide a comprehensive analysis of PE3R’s design, we present key ablation studies here, with additional experiments detailed in the supplementary material. All evaluations are conducted on both open-vocabulary segmentation and multi-view depth estimation to quantitatively assess the contribution of each component.

Effect of Disambiguation Modules. We first analyze the three core modules responsible for establishing a consistent semantic space: Multi-Level Disambiguation, Cross-View Disambiguation, and Global Min-Max Normalization. Quantitative results on ScanNet++ (Table 5) demonstrate that ablating any module causes a performance drop, with the most significant degradation occurring when Multi-Level Disambiguation is removed. This underscores its critical role in resolving hierarchical ambiguities and maintaining part-whole consistency. Qualitative results in Figs. 3–5 offer further insight. As shown in Fig. 3, the absence of Multi-Level Disambiguation leads to fragmented semantics. While parts (*e.g.*, “Drip tray”) may be detected, the model fails to associate them with the whole object (*e.g.*, “Espresso machine”). Our module enables coherent hierarchical understanding. Fig. 4 illustrates that without Cross-View Disambiguation, the same object (*e.g.*, “Bicycle” or “Chair”) is inconsistently labeled across viewpoints due to occlusion or perspective changes. Integrating this module restores semantic coherence throughout the 3D scene. Fig. 5 reveals that removing Global Min-Max Normalization results in uncalibrated similarity scores across views, impairing retrieval reliability. This step ensures a consistent and interpretable similarity range, which is crucial for robust open-vocabulary localization. To validate our specific design choice for feature aggregation, we compare our geometry-aware interpolation against naive uniform averaging and confidence-weighted baselines using DUST3R’s per-point confidence (conf-avg and conf-slerp). As shown in Table 6, uniform averaging leads to consistently lower mIoU and less accurate retrieval, while our area-weighted slerp outperforms both confidence-weighted variants on ScanNet++. This confirms that our area-weighted, spherical interpolation is essential for generating discriminative and reliable semantic embeddings.

Method	rel↓	τ ↑	Runtime	Δ
w/o Semantic Recon.	4.9	64.4	10.40s	-
PE3R (full)	4.5	68.0	11.19s	+0.79s

Table 7. Ablation on Semantic Point Cloud Reconstruction. Removing this module degrades geometric quality.



Figure 6. Effect of Semantic Point Cloud Reconstruction. Removing this module increases spatial noise and reduces object-level coherence in the reconstructed point cloud, highlighting its importance for both geometric and semantic fidelity.

Effect of Semantic Reconstruction Module. We further ablate the Semantic Point Cloud Reconstruction module to evaluate its impact on geometric accuracy. The results in Table 7 show that this module brings a noticeable improvement in depth estimation metrics with only a minimal computational overhead. Qualitatively, Fig. 6 demonstrates that semantic-guided refinement effectively suppresses spatial outliers caused by challenging conditions like reflections and transparency, resulting in cleaner and more structurally coherent point clouds. This confirms that semantic cues actively contribute to enhancing the underlying geometry, beyond merely providing labels.

5. Conclusion

We presented PE3R, a framework that establishes a new paradigm for efficient and generalizable 3D semantic reconstruction from unposed images. By integrating pixel embedding disambiguation, semantic point cloud reconstruction, and global view perception into a cohesive pipeline, PE3R achieves robust zero-shot generalization across diverse scenes without any scene-specific fine-tuning. Extensive evaluations across diverse benchmarks demonstrate that PE3R achieves state-of-the-art accuracy with good efficiency. We believe PE3R paves the way for more scalable and practical 3D vision systems, with promising applications in robotics, AR/VR, and autonomous navigation.

Acknowledgements. This project is supported by the National Research Foundation, Singapore, under its Medium Sized Center for Advanced Robotics Technology Innovation.

References

- [1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjarholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016. 1, 6
- [2] Vladislav Ayzenberg and Marlene Behrmann. Development of visual object recognition. *Nature Reviews Psychology*, 3(2):73–90, 2024. 1
- [3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 1, 6
- [4] Yash Bhalgat, Iro Laina, Joao F Henriques, Andrew Zisserman, and Andrea Vedaldi. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. *arXiv preprint arXiv:2306.04633*, 2023. 2
- [5] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, et al. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems*, 36:25971–25990, 2023. 1, 2
- [6] Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians, 2024. 1, 2
- [7] Hanlin Chen, Chen Li, Mengqi Guo, Zhiwen Yan, and Gim Hee Lee. Gnesf: Generalizable neural semantic fields. *Advances in Neural Information Processing Systems*, 36:36553–36565, 2023. 2
- [8] Seokhun Choi, Hyeonseop Song, Jaechul Kim, Taehyeong Kim, and Hoseok Do. Click-gaussian: Interactive segmentation to any 3d gaussians. *arXiv preprint arXiv:2407.11793*, 2024. 2
- [9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 6
- [10] Bardienu Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. *arXiv preprint arXiv:2409.19152*, 2024. 2
- [11] Francis Engelmann, Fabian Manhardt, Michael Niemeyer, Keisuke Tateno, Marc Pollefeys, and Federico Tombari. OpenNeRF: Open Set 3D Neural Scene Segmentation with Pixel-Wise Features and Rendered Novel Views. In *International Conference on Learning Representations*, 2024. 4
- [12] Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, et al. Large spatial model: End-to-end unposed images to semantic 3d. *Advances in Neural Information Processing Systems*, 37:40212–40229, 2024. 2, 4, 5, 6
- [13] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35:3403–3416, 2022. 2
- [14] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 6
- [15] Rahul Goel, Dhawal Sirikonda, Saurabh Saini, and PJ Narayanan. Interactive segmentation of radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4201–4211, 2023. 1
- [16] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023. 2
- [17] Yansong Guo, Jie Hu, Yansong Qu, and Liujuan Cao. Wildseg3d: Segment any 3d objects in the wild from 2d images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5166–5176, 2025. 1
- [18] Xu Hu, Yuxi Wang, Lue Fan, Junsong Fan, Junran Peng, Zhen Lei, Qing Li, and Zhaoxiang Zhang. Semantic anything in 3d gaussians. *arXiv preprint arXiv:2401.17857*, 2024. 1, 2
- [19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2
- [20] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lurf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 1, 4, 5, 6
- [21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3, 4
- [22] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 1, 6
- [23] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35:23311–23330, 2022. 1
- [24] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 1, 2, 6
- [25] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

- [26] Yichen Liu, Benran Hu, Junkai Huang, Yu-Wing Tai, and Chi-Keung Tang. Instance neural radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 787–796, 2023. 2
- [27] Yichen Liu, Benran Hu, Chi-Keung Tang, and Yu-Wing Tai. Snerf-hq: Segment anything for nerf in high quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3216–3226, 2024. 1
- [28] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*, pages 210–227. Springer, 2022. 2
- [29] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2
- [30] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2
- [31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3
- [32] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–824, 2023. 1
- [33] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 1, 2, 4, 5
- [34] Yansong Qu, Shaohui Dai, Xinyang Li, Jianghang Lin, Lijuan Cao, Shengchuan Zhang, and Rongrong Ji. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5328–5337, 2024. 2, 4, 5, 6
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5
- [36] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 4, 6
- [37] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 6
- [38] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 6
- [39] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 1, 6
- [40] Philipp Schröppel, Jan Bechtold, Artemij Amiranashvili, and Thomas Brox. A benchmark and a baseline for robust multi-view depth estimation. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2022. 6
- [41] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Flashsplat: 2d to 3d gaussian splatting segmentation solved optimally. *arXiv preprint arXiv:2409.08270*, 2024. 2
- [42] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kotschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023. 2
- [43] Pawan Sinha and Tomaso Poggio. Role of learning in three-dimensional form perception. *Nature*, 384(6608):460–463, 1996. 1
- [44] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1, 6
- [45] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. 1
- [46] Songlin Tang, Wenjie Pei, Xin Tao, Tanghui Jia, Guangming Lu, and Yu-Wing Tai. Scene-generalizable interactive segmentation of radiance fields. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6744–6755, 2023.
- [47] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *2022 International Conference on 3D Vision (3DV)*, pages 443–453. IEEE, 2022. 1
- [48] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 2
- [49] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21686–21697, 2024. 2
- [50] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt:

- Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025. [1](#), [2](#), [6](#)
- [51] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. [2](#)
- [52] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. [1](#), [2](#), [3](#), [4](#), [6](#)
- [53] Yiqun Wang, Ivan Skorokhodov, and Peter Wonka. Hf-neus: Improved surface reconstruction using high-frequency details. *Advances in Neural Information Processing Systems*, 35:1966–1978, 2022. [2](#)
- [54] Andrew E Welchman, Arne Deubelius, Verena Conrad, Heinrich H Bülthoff, and Zoe Kourtzi. 3d shape perception from combined depth cues in human visual cortex. *Nature neuroscience*, 8(6):820–827, 2005. [1](#)
- [55] Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. [2](#), [6](#)
- [56] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. *arXiv preprint arXiv:2312.00732*, 2023. [1](#), [2](#), [4](#), [5](#)
- [57] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. [1](#), [6](#)
- [58] Hong-Xing Yu, Leonidas J Guibas, and Jiajun Wu. Unsupervised discovery of object radiance fields. *arXiv preprint arXiv:2107.07905*, 2021. [2](#)
- [59] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. [2](#)
- [60] Chaoning Zhang, Dongshen Han, Sheng Zheng, Jinwoo Choi, Tae-Ho Kim, and Choong Seon Hong. Mobilesamv2: Faster segment anything to everything. *arXiv preprint arXiv:2312.09579*, 2023. [6](#)
- [61] Shuai Feng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. [2](#)
- [62] 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. [1](#), [2](#), [4](#), [5](#)