Unified Semantic Transformer for 3D Scene Understanding

Sebastian Koch[1,2†] Johanna Wald[2] Hide Matsuki[2]
Pedro Hermosilla[3] Timo Ropinski[1] Federico Tombari[2,4]
[1]University Ulm [2]Google [3] TU Vienna [4]TU Munich

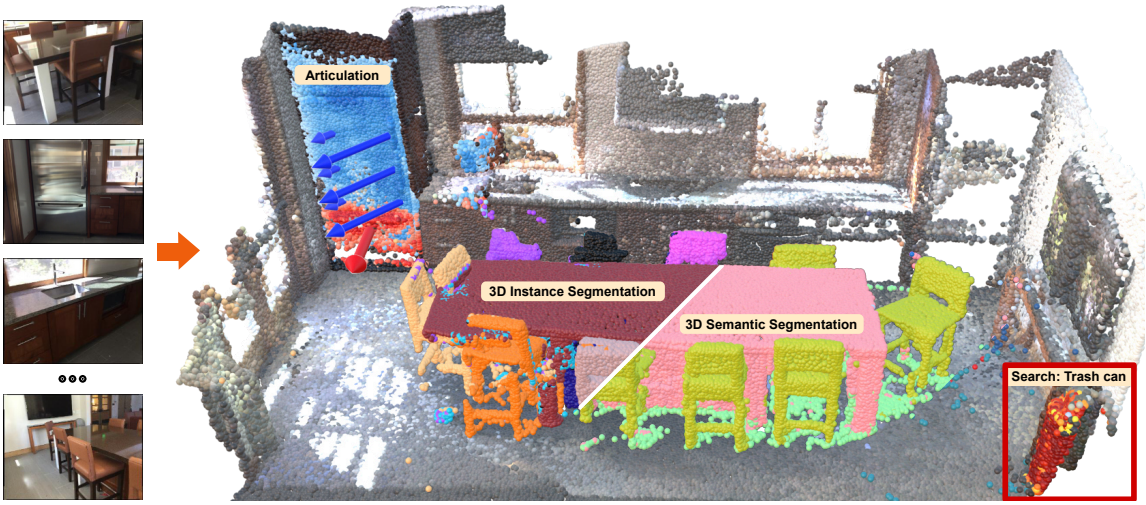

Figure 1. **UNITE Overview.** Given a set of images, UNITE reconstructs both the 3D scene geometry and 3D features used to perform semantic segmentation, instance segmentation, open-vocabulary search and articulation prediction.

Abstract

Holistic 3D scene understanding involves capturing and parsing unstructured 3D environments. Due to the inherent complexity of the real world, existing models have predominantly been developed and limited to be task-specific. We introduce UNITE, a Unified Semantic Transformer for 3D scene understanding, a novel feed-forward neural network that unifies a diverse set of 3D semantic tasks within a single model. Our model operates on unseen scenes in a fully end-to-end manner and only takes a few seconds to infer the full 3D semantic geometry. Our approach is capable of directly predicting multiple semantic attributes, including 3D scene segmentation, instance embeddings, open-vocabulary features, as well as affordance and articulations, solely from RGB images. The method is trained using a combination of 2D distillation, heavily relying on self-supervision and leverages novel multi-view losses designed to ensure 3D view consistency. We demonstrate that UNITE achieves state-of-the-art performance on several different semantic tasks and even outperforms task-specific models, in many cases, surpassing methods that operate on ground truth 3D geometry. See the project website at unite-page.github.io

1. Introduction

3D scene understanding is the foundation of applications in AR/VR and robotics by enabling systems to perceive their surroundings and construct rich 3D representations that combine geometry, such as meshes, point clouds, and TSDF, with high-level semantics of object entities, class semantics, their material, state, or even affordances. While recent advances in foundation models [10, 36, 40, 58] and multi-modal LLMs [5, 27, 28] greatly improved the extraction of semantics from 2D images; these approaches exclusively operate on 2D inputs and therefore do not fully leverage the geometric reasoning available in 3D data.

To transfer the knowledge of strong 2D models into multi-view consistent 3D representations, three main strategies have emerged. First, radiance-field methods [6, 22, 25, 39] learn multi-view features from 2D foundation models alongside NeRF [32] or Gaussian Splatting [21]. They regress consistent 3D features but depend on known camera poses and scene-specific training, which does not generalize to new environments. Second, distillation methods like OpenScene [37] distill 2D features into 3D networks that operate on point clouds directly by aligning their fea-

† This work was conducted during an internship at Google.

ture spaces with respective losses. These methods then need the 3D reconstruction at inference time. Finally, some approaches propose simple but effective lifting-based techniques [12, 46, 47, 54] to decouple 3D geometry and high-level semantics by projecting 2D predictions into an explicit 3D reconstruction. This modular design is non-differentiable and depends on hand-crafted view selection or scene segmentation, often carefully tuned for the specific application domain, which does not scale well. Since view-dependent features of 2D models are sensitive to background noise and limited context, achieving strong multi-view semantic consistency remains challenging.

Recently, transformer-based feed-forward models for 3D reconstruction have demonstrated a breakthrough in multi-view geometric consistency by unifying tasks within a single architecture. Approaches such as DUSt3R [51] and VGGT [50] recover camera poses, depth maps, point maps, and point tracks in a single feed-forward pass from RGB images alone. Despite their success, these methods address only geometric scene attributes and do not model the semantic properties of the input. While several works have explored their extension to semantic 3D tasks, they still rely on explicit 2D-to-3D lifting, preventing the model from learning geometric and semantic representations in a fully unified manner.

This paper addresses these limitations by introducing UNITE, a large feed-forward transformer capable of geometry-grounded semantic prediction. Our method achieves native 3D semantic consistency by jointly learning geometry and semantics within a single network, avoiding any hand-designed lifting steps and enabling a truly end-to-end formulation for diverse semantic tasks. To this end, we propose the following contributions:

- We introduce UNITE, a semantic feed-forward transformer that predicts a full 3D reconstruction with the key 3D semantic attributes from multi-view images, including semantic and instance segmentation, open-vocabulary queries for objects and their affordances, as well as object articulation.
- We train the unified model end-to-end using distillation from 2D foundation models, mostly relying on self-supervision, and enforce a novel multi-view consistency loss for consistent 3D semantic predictions.
- UNITE achieves state-of-the-art results on all semantic tasks and outperforms other semantic feed-forward models as well as other lifting methods, even improving upon methods that operate on ground truth 3D geometry.

2. Related Work

3D Scene Understanding. Early 3D scene understanding approaches were task-specific, closed-vocabulary, and overfitted to benchmarks, addressing specialized problems such as 3D object detection [33, 38], semantic segmentation [34, 55], instance segmentation [13, 43, 48] or affordance and articulation prediction [4, 30] on point clouds and RGB-D images. These methods were typically trained on curated datasets such as ScanNet [3], 3RScan [49], or MultiScan [30], limiting their generalization beyond those domains. Recent work has shifted toward open-vocabulary prediction [6, 22, 24, 37, 39, 46], enabling more flexible deployment across diverse scenes and domains. However, this transition has been largely driven by 2D foundation models [23, 36, 40] trained on internet-scale data, making it more challenging to develop a purely 3D open-vocabulary approach. To this end, many methods have integrated vision-language models such as CLIP [40] or SAM [23], not only during training [6, 24, 37] but also at inference through hand-crafted, non-end-to-end pipelines [12, 35, 46, 47, 54]. This dependency substantially increases the computational and data requirements of 3D scene understanding pipelines, as they often rely on both complete 3D reconstructions and aligned RGB-D images with known poses and intrinsics.

Feed-Forward Models. Classically, image-based dense 3D reconstruction has been approached by integrating multiple subproblems, such as keypoint detection, matching, bundle adjustment, and multi-view stereo. DUSt3R [51] marked a paradigm shift, demonstrating that a single transformer could effectively solve these subproblems just by a minimal post-processing of dense point map prediction. VGGT [50] and subsequent works [20, 29, 53] have further shown that the model can take an arbitrary number of input views and predict diverse geometric attributes.

This success has motivated efforts toward unified 3D semantic scene understanding using a similar end-to-end approach. Initial works addressed the geometric-only limitation by projecting CLIP features into 3D with DUSt3R or VGGT [11, 16]. Other methods introduced feature 3D Gaussian Splatting heads on top of DUSt3R and VGGT to render CLIP-aligned features [8, 45]. Concurrent efforts, integrate SAM features into VGGT but still rely on CLIP feature lifting for open-vocabulary segmentation [26]. While promising, these methods rely on semantic features lifted from 2D foundation models, resulting in a lifting process that often uses hand-designed fusion algorithms [16] and fails to achieve the native, fully learned 3D consistency seen in the original geometric models.

SAB3R [1] addressed this consistency issue to some extent by learning a dedicated semantic head directly on top of the DUSt3R backbone, yet they do not explicitly enforce any multi-view consistency between features. Among the various existing semantic feed-forward model works, PanSt3R [60] might be the closest work to ours, as it achieves multi-view consistent semantic prediction purely through a single feed-forward pass. PanSt3R combines MUSt3R with a dedicated segmentation architecture, Mask2Former [2]. However, its reliance on a separate model

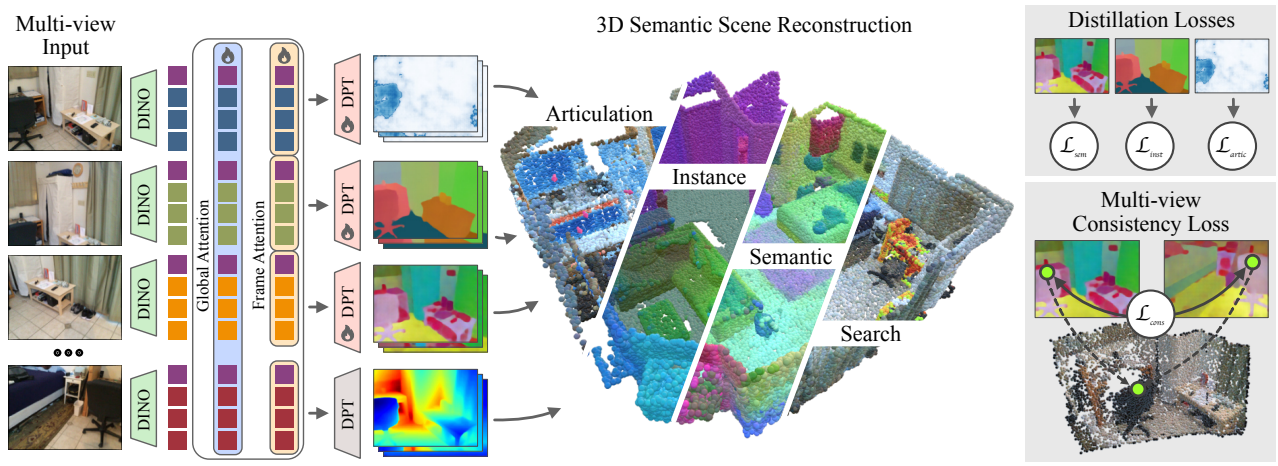

Figure 2. **Method Overview.** Given a set of RGB images, we predict a point cloud reconstruction of the scene with dense features for each point. We train a feed-forward network end-to-end to encode semantic, instance and articulation features, supervised by foundation model features and labels on 2D images. For each output view, we enforce a multi-view consistency loss for corresponding points to ensure that pixels mapping to the same point share the same feature.

architecture introduces a multi-stage dependency and limits the number of input views (typically < 50), which inherently constrains its scalability to larger or more complex scenes.

In contrast, our objective is to develop a single, unified feed-forward model that jointly estimates the geometric structure and multiple semantic attributes for an arbitrary number of input views. Crucially, our approach does not require any hand-designed post-processing or reliance on separate task-specific networks, making it fully trainable in an end-to-end manner for holistic scene understanding.

3. Method

Given a set of RGB images $\mathcal{I} = I_{i\,i=1}^{N}$ capturing a scene from multiple viewpoints, our goal is to reconstruct a holistic 3D representation that provides semantic and instance-level features at different granularities, captures object affordances through open-vocabulary reasoning, and models object articulations alongside per-point geometry, in order to enable interactive querying of concepts and functionalities within the given scene.

To this end, we propose an end-to-end trainable multi-task network F_θ that jointly infers the 3D structure and its semantic and functional attributes in a single forward pass:

$$F_\theta\big(\{I_i\}_{i=1}^{N}\big) \;\mapsto\; \big\{k_i,\, D_i,\, P_i,\, S_i,\, G_i,\, A_i\big\}_{i=1}^{N}, \qquad (1)$$

where k_i denotes the camera intrinsics and pose, D_i the predicted depth map, P_i the point map representing per-pixel 3D coordinates, S_i the semantic features (Sec. 3.2), G_i the instance grouping features (Sec. 3.3), and A_i the articulation predictions for each image (Sec. 3.4). All tasks are trained jointly using a combination of 2D and 3D supervision signals, complemented by a multi-view consistency objective that enforces agreement across views observing the same 3D point. An overview of the proposed framework, UNITE, is shown in Fig. 2.

3.1. Geometric Foundation

To extract the geometric attributes of the scene, we build upon recent pre-trained feed-forward models for geometry prediction, specifically VGGT [50], as a foundation. A set of multi-view images is encoded using a pre-trained image encoder [36], and the resulting image tokens are processed through k blocks of alternating frame-wise and global self-attention. This backbone integrates information across views and, through dedicated heads, predicts camera poses and point maps, providing a strong geometric foundation through its large-scale pre-training and multi-view reasoning capability. The resulting representations offer a promising basis for semantic understanding, as they capture spatial consistency and geometric structure across views.

3.2. Semantics Features

We propose a Dense Prediction Transformer (DPT) [41] which learns open-vocabulary semantics on top of the shared multi-view fusion backbone. This design ensures that, in a multi-task setting, each head can learn task-specific features while sharing the strong multi-view fusion backbone. Thus, we are able to predict multi-view consistent features in CLIP space, denoted as $f_{\text{sem}} \in \mathbf{R}^{W\times H\times d_s}$, along with a per-pixel semantic confidence score $c_{\text{sem}} \in [1, \infty)$.

Distillation. We supervise the DPT-Head using dense 2D features $f_{\text{sem}}^{2D} \in \mathbf{R}^{W\times H\times d}$, extracted from the corresponding RGB images. To obtain these features, we first segment each image using SAM [23] and then encode each segment using CLIP, which serves as the vision-language feature extractor. The semantic distillation loss is defined as a cosine similarity loss between the predicted feed-forward features

and the CLIP-encoded 2D features:

$$\mathcal{L}_{\text{sem}}^{2D} = 1 - \cos(f_{\text{sem}}, \hat{f_{\text{sem}}^{2D}}), \tag{2}$$

which encourages the predicted features to be co-located in the vision-language representation.

Multi-View Consistency. We find that the 2D model provides view-inconsistent features for the same object when observed from different viewpoints, causing inconsistencies that produce a contradictory training signal, which hampers our method's goal of predicting multi-view consistent 3D semantic reconstruction. To address this issue, we introduce a multi-view consistency loss that enforces feature agreement across different projections of the same 3D point. Let p denote a 3D point in the point cloud $\mathcal{Q}$, and $\mathcal{I}_p \subseteq \mathcal{I}$ the set of views from which this point is visible. Using the camera poses, intrinsics, and depth maps, we determine pixel correspondences $\text{corr}(p) = \{u_i \mid p \text{ is visible in } I_i \text{ at pixel } u_i\}$, where $u_i \in \Omega_i$ denotes the image coordinates in view I_i of point p. For each correspondence, we compute a confidence-weighted mean feature

$$\bar{f}_p = \frac{\sum_{u_i \in \text{corr}(p)} c_{p,u_i} \, f_{p,u_i}}{\sum_{u_i \in \text{corr}(p)} c_{p,u_i}}, \tag{3}$$

where f_{p,u_i} and c_{p,u_i} denote the predicted feature vector and confidence for point p observed in view I_i at pixel u_i, respectively. This weighted aggregation enables the model to learn which views provide more reliable features, while enforcing consistent semantics across all views of the same 3D point. For each query point $p \in \mathcal{Q}$, we first define a per-point consistency term

$$\ell_p^{\text{cons}} := \frac{1}{|\text{corr}(p)|} \sum_{u_i \in \text{corr}(p)} \Big[1 - \cos\big(f_{p,u_i}, \, \text{stopgrad}(\bar{f}_p)\big)\Big], \tag{4}$$

which measures the alignment between view-specific features and the aggregated feature $\bar{f}_p$, while the stop-gradient operator $\text{stopgrad}(\cdot)$ prevents gradients from flowing into $\bar{f}_p$. The overall multi-view consistency loss is then given by

$$\mathcal{L}_{\text{cons}} = \frac{1}{|\mathcal{Q}|} \sum_{p \in \mathcal{Q}} \ell_p^{\text{cons}}, \tag{5}$$

which encourages feature consistency across views and ensures that the learned representations remain view-invariant.

The overall semantic objective combines the distillation and consistency terms as

$$\mathcal{L}_{\text{sem}} = \lambda_{\text{sem}}^{2D} \, \mathcal{L}_{\text{sem}}^{2D} + \lambda_{\text{cons}} \, \mathcal{L}_{\text{cons}}^{\text{sem}}, \tag{6}$$

where λ_{sem2D} and λ_{cons} are weighting coefficients controlling the balance between 2D semantic alignment and multi-view feature consistency. This objective encourages the model to learn open-vocabulary semantic features aligned with the vision-language embedding space and consistent across multiple views of the same 3D point.

3.3. Instance Features

Rather than employing conventional mask-prediction networks such as Mask R-CNN [15] or Mask2Former [2], which infer discrete per-view instance masks, we adopt a contrastive formulation that learns a metric embedding space. In this space, features of pixels belonging to the same instance are pulled together, while those from different instances are pushed apart. This design naturally supports multi-view consistency, as instance identities are aligned through feature similarity instead of explicit mask matching, enabling consistent supervision across viewpoints. To this end, following the same approach as the open-vocabulary semantic head, we predict instance features $g_{\text{inst}} \in \mathbb{R}^{W \times H \times d_g}$ through a DPT head on top of the shared multi-view backbone. This simple extension allows for instance, reasoning to leverage the fused multi-view context, while contributing complementary gradients that strengthen the shared encoder. Similar to the open-vocabulary head, this head is supervised by a 2D foundation model, requiring no manual annotations and scaling easily to large datasets. Specifically, we utilize SAM [23] to extract class-agnostic instance segmentation masks from the RGB input images.

To ensure multi-view consistent supervision, the 2D instance masks are first lifted into 3D space using ground-truth depth and camera parameters, and grouped using DBSCAN [7] to produce 3D consistent masks. These masks are then projected back into all views in which they are visible, enabling cross-view supervision with consistent instance correspondences. This 3D-aware distillation ensures that pixels corresponding to the same physical instance, even when observed from different viewpoints, are encoded with consistent instance features.

We formulate a pairwise contrastive loss in feature space to train the dense instance embeddings g_{inst}. Given two pixels u_i and u_j from views with corresponding instance assignments l_{u_i} and l_{u_j} provided by SAM, we define the instance contrastive loss as

$$\mathcal{L}_{\text{grouping}}^{2D} = \begin{cases} \|g_{u_i} - g_{u_j}\|_2, & l_{u_i} = l_{u_j}, \\ \text{ReLU}[m - \|g_{u_i} - g_{u_j}\|_2], & l_{u_i} \neq l_{u_j}, \end{cases} \tag{7}$$

where g_{u_i} denotes the instance feature at pixel u_i, and m is a margin that enforces a minimum separation between embeddings of different instances.

To further promote view-invariant instance representations, we apply the generalized multi-view consistency formulation introduced in Eq. (5), reusing the same confidence-weighted aggregation from Eq. (3) for the instance features g_{inst}. The final instance objective combines the grouping and consistency terms as

$$\mathcal{L}_{\text{inst}} = \lambda_{\text{group}} \, \mathcal{L}_{\text{grouping}}^{2D} + \lambda_{\text{cons}} \, \mathcal{L}_{\text{cons}}^{\text{inst}}, \tag{8}$$

where λ_{group} and λ_{cons} control the balance between intra-instance compactness and cross-view consistency. This

formulation encourages the model to learn dense, class-agnostic instance features that cluster tightly within instances, remain well-separated across different instances, and stay consistent across all viewpoints.

3.4. Articulations

To extend scene understanding beyond static geometry and semantics, we additionally predict object articulations in a dense, per-pixel manner, also by using a dedicated DPT head. Articulations capture how object parts can move and provide functional information essential for reasoning about interactions within a scene.

We consider two fundamental articulation types, that we learn from annotated data in 3D: translational motions, such as drawers sliding along a linear path, and rotational motions, such as doors rotating around a hinge. To represent both within a single regression space, we approximate rotational motion as linear ones. This approximation is computed from ground-truth articulation parameters: for each object with a known rotation axis, we rotate its surface points by $90\,\mathrm{deg}$ around this axis and use the displacement between original and rotated positions as the linearized ground-truth motion direction.

Because articulated objects are sparse in typical scenes, we decompose the prediction into two tasks: (i) identifying which points belong to articulated objects, and (ii) predicting their local motion directions. The articulation head outputs a per-pixel articulation probability map $p_{\text{pred}} \in [0,1]^{W\times H}$ with a confidence map c_{exist}, and a per-pixel 3D *vector map* $\mathbf{v}_{\text{pred}} \in \mathbb{R}^{W\times H\times 3}$ with a corresponding confidence map c_{motion}. These vector maps encode the articulation motion at each pixel, analogous to how point maps [51] assign a 3D coordinate to each pixel.

Articulation existence is trained using a binary ground-truth map $\hat{p}$ with a focal loss to address class imbalance:

$$\mathcal{L}_{\text{exist}} = \frac{1}{M}\sum_{i=1}^{M} \mathcal{L}_{\text{Focal}}\big(p_i, \hat{p}_i\big). \tag{9}$$

For pixels corresponding to articulated regions, motion vectors $\mathbf{v}_i$ are regressed toward ground-truth vectors $\hat{\mathbf{v}}_i$ using an ℓ_2 loss:

$$\mathcal{L}_{\text{motion}} = \frac{1}{M_A}\sum_{i\in M_A} \|\mathbf{v}_i - \hat{\mathbf{v}}_i\|_2^2, \tag{10}$$

where M_A is the set of articulated pixels.

To enforce correspondence across viewpoints, we apply the confidence-weighted multi-view consistency from Eqs. (3) and (5) to both the existence and vector-map. The full articulation loss thus becomes:

$$\begin{aligned}\mathcal{L}_{\text{artic}} = \lambda_{\text{exist}}\,\mathcal{L}_{\text{exist}} + \lambda_{\text{cons}}^{\text{exist}}\,\mathcal{L}_{\text{cons}}^{\text{exist}} \\ + \lambda_{\text{motion}}\,\mathcal{L}_{\text{motion}} + \lambda_{\text{cons}}^{\text{motion}}\,\mathcal{L}_{\text{cons}}^{\text{motion}}.\end{aligned} \tag{11}$$

3.5. Multi-Task Optimization

Finally, the network is optimized end-to-end using a weighted combination of all task-specific objectives, enabling shared features to benefit from complementary supervision across tasks

$$\mathcal{L} = \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{inst}}\mathcal{L}_{\text{inst}} + \lambda_{\text{artic}}\mathcal{L}_{\text{artic}}.^{1} \tag{12}$$

4. Experiments

In this section, we evaluate our approach across three open-vocabulary 3D scene understanding tasks — semantic segmentation, instance segmentation, and articulation prediction — on ScanNet [3], ScanNet200 [42], ScanNet++ [57], and MultiScan [30]. Our method achieves state-of-the-art performance on all tasks, consistently surpassing task-specific baselines and demonstrating the benefits of a unified multi-task network.

Unlike prior work that reports view-level or 2D metrics, all evaluations in this paper are performed directly in 3D to assess genuine 3D scene understanding in semantic feed-forward models. Sec. 4.1 presents quantitative results for open-vocabulary 3D semantic segmentation, followed by class-agnostic instance segmentation in Sec. 4.2. In Sec. 4.3, we show the versatility of our model for articulation and affordance prediction. Sec. 4.4 provides qualitative results demonstrating generalization to diverse indoor environments, and Sec. 4.5 analyzes the impact of multi-view consistency, feature aggregation, and the joint geometric-semantic formulation.

4.1. 3D Semantic Segmentation

Setup. We evaluate our method on *ScanNet* [3], *ScanNet200* [42], and *ScanNet++* [57] for 3D semantic segmentation. These datasets differ in the number and granularity of semantic classes, and we evaluate performance on all of them using standard class-wise mIoU and mAcc on the ground truth 3D point clouds.

We compare our approach against recent SLAM-based open-vocabulary methods [12, 54], which jointly reconstruct the scene and extract open-vocabulary features, operating under privileged conditions with access to ground-truth depth, intrinsics, and camera poses. We also include recent semantic feed-forward approaches [8, 16, 45, 60] using their public implementations. For fairness, all methods receive up to 200 images per scene. To isolate semantic performance and ensure comparability with SLAM-based methods, we additionally provide feed-forward models with ground-truth camera parameters and poses. Their image-based frustum predictions are aligned with the ground-truth point cloud using a one-nearest-neighbor mapping. If a feed-forward model cannot process all 200 images at once [8, 16], we split

[1] A full overview of the loss weights is given in the supplementary.

the sequence into smaller subsets and align each subset's predictions using the corresponding ground-truth poses.

Finally, we compare against established point cloud-based open-vocabulary methods, which are the most privileged, as they operate on posed RGB-D images and have access to the evaluation meshes at inference time. We omit comparisons with open-vocabulary radiance field approaches [6, 22, 39] since they require scene-specific training, unlike our method and the baselines.

Class prediction. The 3D semantic segmentation is obtained by querying the 3D semantic feature representation with the benchmark set of text labels and assigning the class with the highest similarity score to each 3D point. Querying is performed by computing the cosine similarity between the point cloud features and the text embeddings of the vision language model, specifically here the CLIP-text encoder.

Results. Tab. 1 compares UNITE against our point cloud-based, SLAM-based and feed-forward model-based baselines. UNITE outperforms all SLAM and semantic forward-model baselines by a large margin for both mIoU and mAcc. Our closest competitor PanSt3R [60], performs well on ScanNet++ [57] whose labels were contained in the Mask2Former [2] training of PanSt3R [60], however on ScanNet20 and ScanNet200 our approach outperforms PanSt3R by a large margin. Furthermore, our approach is the only RGB(-D) approach that outperforms a native point cloud-based 3D scene understanding approach in OpenScene [37]. However, we observe a performance gap compared to the current point cloud-based state-of-the-art OV3D on ScanNet20. We attribute this to OV3D's many optimizations combining vision and large language models which seem to help in small label scenarios but hurt performance on larger label sets such as ScanNet200 where we outperform OV3D.

4.2. 3D Instance Segmentation

Setup. Instance segmentation is evaluated on *ScanNet* [3], *ScanNet200* [42], and *ScanNet++* [57], comparing our approach to representative unsupervised, lifting-based, and feed-forward methods. We omit comparisons with approaches trained on instance mask annotations, such as Mask3D [43], since these methods rely on dense supervision unavailable in our setting, making such comparisons unfair and not representative of the unsupervised or weakly supervised nature of our approach. Following the ScanNet evaluation protocol, we report Average Precision (AP) at mask overlap thresholds of 50% and 25%, as well as the mean AP averaged over IoU thresholds from 0.50 to 0.95 in steps of 0.05. Because our model performs class-agnostic instance segmentation, we ignore semantic class labels when matching predictions to the ground-truth instances.

We compare against several class-agnostic baselines:

Table 1. **3D Semantic Segmentation.** We compare our method with open-vocabulary 3D point cloud methods, SLAM approaches and semantic feed-forward models evaluated with the ground truth 3D segmentation. † For fair comparison, we evaluate Panst3R without their introduced QUBO post-processing.

	ScanNet [3]		*ScanNet200* [42]		*ScanNet++* [57]	
	mIoU	**mAcc**	**mIoU**	**mAcc**	**mIoU**	**mAcc**
3D point cloud methods						
CLIP-FO3D [59]	30.2	49.1	-	-	-	-
OpenScene 2D [10, 37]	41.4	63.6	-	-	-	-
OpenScene 3D [37]	46.0	66.3	07.3	14.5	-	-
OV3D [19]	**57.3**	**72.9**	**08.7**	**15.2**	-	-
RGB-D methods						
Concept-Graphs [12]	17.1	29.1	06.0	11.7	-	-
HOV-SG [54]	**34.4**	**51.1**	**11.2**	**18.7**	-	-
RGB methods						
Pe3R [16]	10.7	19.8	02.5	06.5	08.3	16.0
Uni3R [45]	29.3	39.4	04.1	06.8	05.2	10.8
LSM [8]	32.2	41.1	06.3	09.7	14.3	26.5
PanSt3R† [60]	42.6	49.7	13.3	19.9	**21.6**	31.4
Ours	**48.7**	**68.3**	**14.5**	**26.3**	17.2	**37.0**

Table 2. **Class-agnostic 3D Instance Segmentation.** We compare our method against different 3D instance segmentation approaches † For fair comparison, we evaluate Panst3R without their introduced QUBO post-processing. Please note, that unlike the other approaches, PanSt3R requires instances labels for training.

	ScanNet			*ScanNet200*			*ScanNet++*		
	AP	**AP$_{50}$**	**AP$_{25}$**	**AP**	**AP$_{50}$**	**AP$_{25}$**	**AP**	**AP$_{50}$**	**AP$_{25}$**
HDBSCAN [31]	1.6	5.5	32.1	2.9	08.2	33.1	4.3	10.6	32.3
Felzenszwalb [9]	5.3	12.6	36.9	04.8	09.8	27.5	**08.8**	**16.9**	36.1
SAM3D [56]	6.3	17.9	47.3	12.1	28.6	54.1	03.0	7.9	22.3
PanSt3R † [60]	11.4	29.3	53.4	10.6	27.3	50.8	06.5	15.9	33.9
Ours	**13.2**	**29.6**	**57.2**	**12.3**	**28.8**	**58.2**	06.6	16.1	**40.4**

HDBSCAN [31], which clusters instances purely from geometry; Felzenszwalb et al. [9], which applies a graph-based segmentation approach; SAM3D [56], which lifts 2D SAM [23] masks into 3D using ground-truth depth and camera parameters; and PanSt3R [60], which employs a Mask2Former [2] head for instance prediction.

Instance Feature Clustering. Our method produces dense instance features, which we cluster to obtain the final class-agnostic 3D instance segmentation. When the number of instances is unknown, we apply HDBSCAN [31]. When the number of instances is known or predefined, we instead use KMeans++ [14].

Results. Tab. 2 reports our class-agnostic 3D instance segmentation results. UNITE achieves state-of-the-art performance on ScanNet and ScanNet200, surpassing other unsupervised approaches, including SAM3D [56], which lifts SAM predictions into 3D using depth maps and merges them via IoU heuristics. Although, UNITE is also trained with SAM masks, our multi-view consistency losses enable significantly stronger 3D-consistent predictions. Moreover, UNITE outperforms the feed-forward baseline PanSt3R [60], the only method trained directly on instance labels. On ScanNet++, however, the classical method of Felzenszwalb et al. [9] performs best, even surpassing our approach on

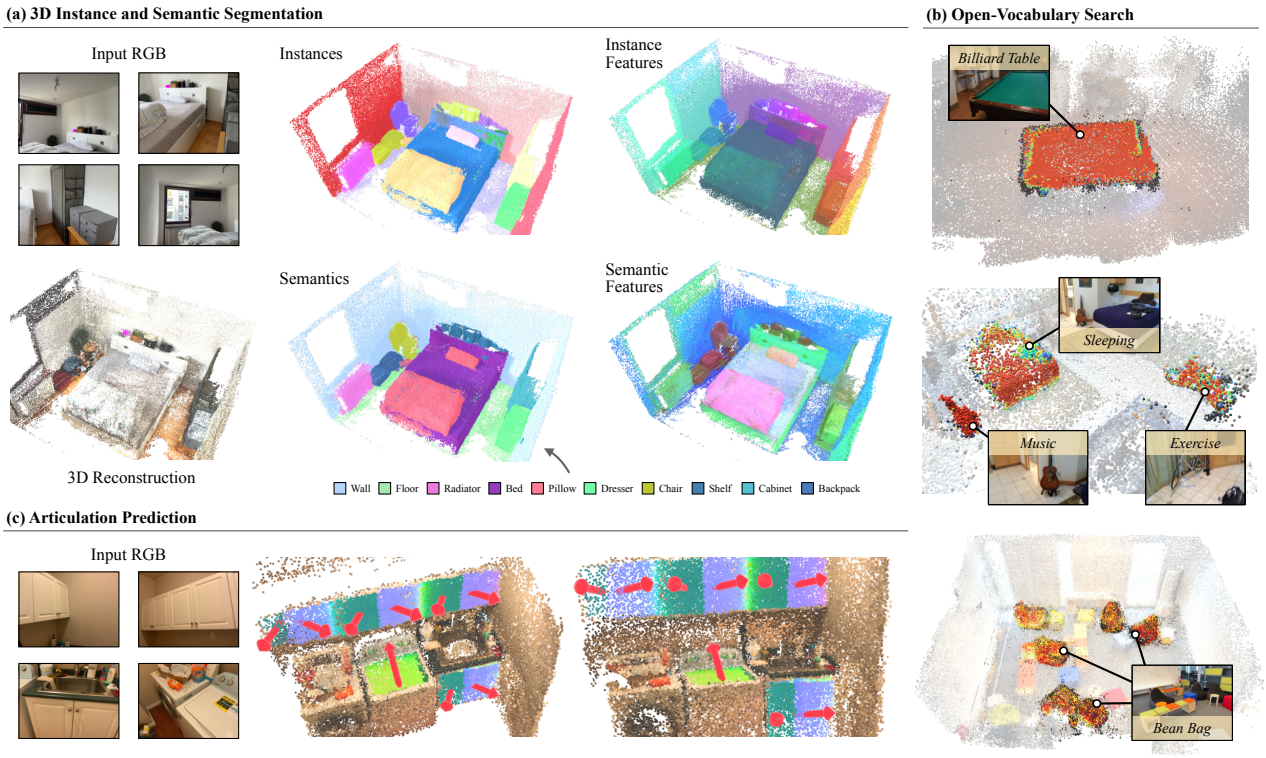

Figure 3. **Results of UNITE on 5 scenes in ScanNet, ScanNet++, and MultiScan.** Each row shows qualitative results a different output modality: (a) 3D semantic and instance segmentation, (b) open-vocabulary search and (c) articulation prediction.

Table 3. **3D Object Articulation Prediction.** We compare against the point cloud-based articulation methods on MultiScan.

		Movable Part			*Motion type*		
	IoU	**R**	**P**	**F1**	**R**	**P**	**F1**
OPDPN [18]	53.8	01.6	03.1	02.1	01.3	02.6	01.7
S2M [52]	69.2	**06.7**	15.7	09.4	**04.5**	10.4	06.3
Ours (task-only)	68.3	05.2	23.4	07.6	01.8	09.2	04.9
Ours (multi-task)	**70.3**	06.0	**29.6**	**10.0**	03.7	**12.3**	**06.9**

AP and AP_{50}. We attribute this to (i) their direct operation on the evaluation mesh, avoiding lifting errors, and (ii) their advantage on fine-grained boundaries, while UNITE yields better AP_{25}, reflecting stronger instance-level recognition.

4.3. Articulation

Finally, we evaluate our method on affordance detection and articulation prediction on the MultiScan [30] dataset. The MultiScan dataset is an extensive dataset that consists of multiple rescans of the same scene to identify articulated objects. Unlike benchmarks such as SceneFun3D [4], which mainly focus on small, affordable elements such as buttons, knobs, and handles, MultiScan focuses on a more holistic understanding of articulated objects with a greater variety, such as drawers, cabinets, rotatable chairs, curtains, etc. This makes MultiScan ideal to evaluate our semantic feed-forward model to test the reconstruction quality and semantic understanding.

Setup. We evaluate UNITE against strong baselines [18, 52] from the MultiScan benchmark. Unlike our approach, these methods operate on reconstructed meshes and predict articulated object parts and their motion type (translation or rotation) in two stages: instance segmentation followed by part segmentation. In contrast, UNITE uses only input images and directly predicts articulated object parts in 3D.

We report IoU, measuring the overlap between our articulation existence prediction and ground-truth articulated parts. In addition, we compute precision, recall, and F1 for movable part detection, considering a part correctly detected if its IoU with the ground truth exceeds 0.5. Finally, we evaluate motion type prediction (translation vs. rotation) under the same IoU threshold to assess the correctness of the predicted articulation type.

Part Articulation Aggregation. UNITE outputs per-pixel articulation vectors together with an articulation-existence probability. To obtain a single articulation prediction for each object or object part, we first discard all motion vectors whose existence probability is below 0.5. The remaining vectors for each predicted instance are then averaged. This yields one unified articulation estimate for every instance segment as visualized in Fig. 3c.

Results. Overall, UNITE outperforms the 3D point cloud baselines on articulation prediction, as shown in Tab. 3. While existing methods rely on reconstructed meshes

and explicit part segmentation, our image-based approach achieves higher IoU and consistently better F1 scores for both movable part and motion type prediction. Notably, our multi-task design, which jointly learns class and instance semantics together with articulation, leads to substantial gains over the task-specific variant. This highlights the benefit of integrating related scene understanding objectives within a single end-to-end feed-forward network, which underlines the benefits of unified, multi-task 3D scene understanding.

4.4. Qualitative Results

In Fig. 3, we show predictions of our approach for 3D semantic and instance segmentation, open-vocabulary instance search, and articulation prediction across scenes from ScanNet [3], ScanNet++ [57], and MultiScan [30]. In Fig. 3a, we perform class-agnostic 3D instance segmentation by clustering the predicted instance features (right) using HDBSCAN [31]. We then assign semantic labels from the ScanNet200 label set by computing cosine similarity between the predicted semantic embeddings and the encoded text label embeddings. The resulting segments exhibit clean object boundaries with minimal noise and accurate semantic predictions, demonstrating geometrically and semantically coherent 3D outputs.

In Fig. 3b, we conduct open-vocabulary instance search using CLIP-aligned features to localize both concrete objects (e.g., a billiard table, bean bags) and higher-level concepts such as sleeping, music, and exercise.

In Fig. 3c, we predict object articulations in a laundry room, correctly identifying movable cabinets and the dryer door along with their estimated motion directions.

4.5. Ablations

We perform ablation studies on three components of our model: joint training versus independent geometric and semantic models, the effect of our multi-view consistency loss with confidence-weighted fusion, and an oracle-geometry upper bound using ground-truth depth and pose. We investigate the following questions:

How does distillation compare to feature reprojection? In this paper, we propose to distill vision-language features from CLIP [40] into a feed-forward model that jointly reconstructs the scene and predicts its semantics. Alternatively, one could combine a feed-forward model such as VGGT [50] with a semantic foundation model like CLIP by lifting CLIP predictions into 3D using the depth and camera parameters predicted by VGGT. However, as shown in Tab. 4, this simple lifting approach performs notably worse than our end-to-end model, which natively performs multi-view fusion. Our approach produces more robust predictions by jointly reasoning across views, whereas the single-view lifting approach is limited by viewpoint occlusions and can only aggregate information through mean pooling without comprehensive multi-view consistency.

Table 4. **Ablations.** We compare a naïve CLIP+VGGT baseline, against our unified geometry–semantics model on ScanNet20. We incrementally add multi-view consistency loss, semantic confidence weighting, and ground truth geometry.

	mIoU	mAcc
VGGT + CLIP Lifting	34.5	46.1
Ours (distillation-only)	44.3	61.9
+ Multi-View consistency (Eq. (5))	45.2 (+0.9)	63.2 (+1.3)
+ Confidence Weight (Eq. (3))	46.3 (+1.1)	64.7 (+1.5)
+ Geometry Oracle	**48.7** (+2.5)	**68.3** (+3.7)

Does semantic confidence improve multi-view fusion? In Eq. (3), we introduce a method to learn multi-view consistent features via a confidence-weighted average of all predicted view features. The per-view confidence encourages the model to rely more on views with clear visibility, while features from occluded or uncertain views are pulled toward those from more confident perspectives. As shown in Tab. 4, this weighted averaging leads to a significant improvement over standard mean pooling. While mean pooling produces a suboptimal embedding that treats all views equally, our confidence-weighted fusion yields a more stable training objective and better overall results.

What if geometry were perfect? When evaluating 3D semantic or instance segmentation, the quality of the geometric reconstruction directly affects the semantic matching, as correspondences are established via nearest-neighbor search between predicted and ground-truth points. To disentangle these factors, we report an upper bound of our method in Tab. 4, obtained by using ground-truth depth and camera parameters. The results show a modest but consistent improvement, indicating that our approach already produces geometry of sufficiently high quality for reliable semantic reasoning.

5. Conclusion

In summary, we introduced UNITE, a unified transformer architecture that achieves 3D semantic scene understanding directly from RGB images by distilling rich 2D foundation model features and enforcing multi-view consistency. While state-of-the-art on multiple tasks, the model's performance is currently tightly coupled to the quality of the underlying geometry. However, rapid advancements in geometric foundation models are directly transferable, with promising immediate improvements of the semantic features.

Also, while we showed the generalizability of the unified model, we only scratched the surface of what we believe is possible. Given the scalability of our training pipeline, the model has the potential of being adapted to larger datasets and expanding to more tasks like scene captioning and decomposition or 4D dynamic segmentation. We believe this opens opportunities for in-the-wild, scalable 3D scene analysis in a truly holistic and unified manner.

References

[1] Xuweiyi Chen, Tian Xia, Sihan Xu, Jianing Yang, Joyce Chai, and Zezhou Cheng. Sab3r: Semantic-augmented backbone in 3d reconstruction. *arXiv preprint arXiv:2506.02112*, 2025. 2

[2] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022. 2, 4, 6

[3] 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. 2, 5, 6, 8, 3

[4] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14531–14542, 2024. 2, 7

[5] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024. 1

[6] 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. 1, 2, 6

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 4

[8] 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, 5, 6

[9] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004. 6

[10] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pages 540–557. Springer, 2022. 1, 6

[11] Ziren Gong, Xiaohan Li, Fabio Tosi, Jiawei Han, Stefano Mattoccia, Jianfei Cai, and Matteo Poggi. Ov3r: Open-vocabulary semantic 3d reconstruction from rgb videos. *arXiv preprint arXiv:2507.22052*, 2025. 2

[12] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024. 2, 5, 6, 3

[13] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2940–2949, 2020. 2

[14] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979. 6

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[16] Jie Hu, Shizun Wang, and Xinchao Wang. Pe3r: Perception-efficient 3d reconstruction. *arXiv preprint arXiv:2503.07507*, 2025. 2, 5, 6

[17] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *Robotics: Science and Systems (RSS)*, 2023. 3

[18] Hanxiao Jiang, Yongsen Mao, Manolis Savva, and Angel X Chang. Opd: Single-view 3d openable part detection. In *European Conference on Computer Vision*, pages 410–426. Springer, 2022. 7

[19] Li Jiang, Shaoshuai Shi, and Bernt Schiele. Open-vocabulary 3d semantic segmentation with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21284–21294, 2024. 6

[20] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. *arXiv:2509.13414*, 2025. 2

[21] 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. 1

[22] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 6, 3

[23] 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, 6, 1

[24] Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[25] Sebastian Koch, Johanna Wald, Mirco Colosi, Narunas Vaskevicius, Pedro Hermosilla, Federico Tombari, and Timo Ropinski. Relationfield: Relate anything in radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1

[26] Hao Li, Zhengyu Zou, Fangfu Liu, Xuanyang Zhang, Fangzhou Hong, Yukang Cao, Yushi Lan, Manyuan Zhang, Gang Yu, Dingwen Zhang, and Ziwei Liu. Iggt: Instance-grounded geometry transformer for semantic 3d reconstruction. *arXiv preprint arXiv:2510.22706*, 2024. 2

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[29] Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025. 2

[30] Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel Chang, and Manolis Savva. Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. *Advances in neural information processing systems*, 35:9058–9071, 2022. 2, 5, 7, 8

[31] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pages 33–42. IEEE, 2017. 6, 8

[32] 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

[33] Ishan Misra, Rohit Girdhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. In *ICCV*, 2021. 2

[34] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3d: Out-of-context data augmentation for 3d scenes. In *2021 international conference on 3d vision (3dv)*, pages 116–125. IEEE, 2021. 2

[35] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[36] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1, 2, 3

[37] 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. 1, 2, 6, 3

[38] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[39] 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, 6

[40] 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. 1, 2, 8

[41] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3

[42] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European conference on computer vision*, pages 125–141. Springer, 2022. 5, 6

[43] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In *International Conference on Robotics and Automation (ICRA)*, 2023. 2, 6

[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. 3

[45] Xiangyu Sun, Haoyi Jiang, Liu Liu, Seungtae Nam, Gyeongjin Kang, Xinjie Wang, Wei Sui, Zhizhong Su, Wenyu Liu, Xinggang Wang, and Eunbyung Park. Uni3r: Unified 3d reconstruction and semantic understanding via generalizable gaussian splatting from unposed multi-view images, 2025. 2, 5, 6

[46] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[47] Ayca Takmaz, Alexandros Delitzas, Robert W Sumner, Francis Engelmann, Johanna Wald, and Federico Tombari. Search3d: Hierarchical open-vocabulary 3d segmentation. *IEEE Robotics and Automation Letters*, 2025. 2

[48] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2708–2717, 2022. 2

[49] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-

localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019. 2

[50] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3, 8, 1

[51] 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 (CVPR)*, pages 20697–20709, 2024. 2, 5

[52] Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinping Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8876–8884, 2019. 7

[53] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning, 2025. 2

[54] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. *Robotics: Science and Systems*, 2024. 2, 5, 6, 3

[55] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024. 2

[56] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 6, 1

[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. 5, 6, 8, 3

[58] 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 (ICCV)*, pages 11975–11986, 2023. 1

[59] Junbo Zhang, Runpei Dong, and Kaisheng Ma. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2048–2059, 2023. 6

[60] Lojze Zust, Yohann Cabon, Juliette Marrie, Leonid Antsfeld, Boris Chidlovskii, Jerome Revaud, and Gabriela Csurka. Panst3r: Multi-view consistent panoptic segmentation. In *International Conference on Computer Vision (ICCV)*, 2025. 2, 5, 6

Unified Semantic Transformer for 3D Scene Understanding

Supplementary Material

In this **supplementary material**, we first provide additional implementation and evaluation details in Sec. A. Next, we report compute requirements and inference-time statistics in Sec. B. We then examine the interaction between geometry and semantics through an ablation study in Sec. C. Finally, in Sec. D, we present out-of-distribution evaluations, including results on Replica and qualitative predictions on LERF scenes. Additionally, we provide a supplementary video attached with the submission.

A. Additional Details

Implementation Details. Our model is built on top of VGGT-1B [50] and is trained on sequences of 12 images. To obtain semantic features, we employ SAM [23] to generate class-agnostic segments for each image. We then extract vision–language features by running OpenSeg [10] over the full image and averaging the resulting features within each SAM segment. This produces a dense feature map of shape $H \times W \times d_s$ with $H = 378, W = 448$ and $d_s = 718$.

For instance feature supervision, we again rely on SAM to extract 2D segments. These segments are lifted into 3D using ground-truth camera parameters and depth maps. We compute 3D IoU between lifted segments and merge them following the procedure of SAM3D [56]. The merged 3D segments are then projected back to the image plane, yielding multi-view consistent mask identifiers for each image. During training, the contrastive margin is set to $m = 1$, and the instance embeddings have dimensionality $d_g = 512$ to ensure sufficient separability in large-scale scenes. At inference time, instance features are clustered using HDBSCAN with eps $= 0.1$.

Articulations are represented using a 9-dimensional vector that includes three translational and three rotational components, a translation-existence mask, a rotation-existence mask, and a prediction confidence value.

The final multi-task objective assigns equal weight to the three main tasks, with $\lambda_{\text{sem}} = 1, \lambda_{\text{inst}} = 1$, and $\lambda_{\text{artic}} = 1$. For each task, we apply a 10-to-1 weighting ratio between the distillation/supervision loss and the multi-view consistency loss. This yields $\lambda_{\text{sem}}^{2D} = 1$ and $\lambda_{\text{cons}} = 0.1$ for semantic feature learning; $\lambda_{\text{group}} = 1$ and $\lambda_{\text{cons}} = 0.1$ for instance feature learning; and $\lambda_{\text{exist}} = 10$, $\lambda_{\text{cons}}^{\text{exist}} = 1$, $\lambda_{\text{motion}} = 1$, and $\lambda_{\text{cons}}^{\text{motion}} = 0.1$ for articulation learning. The model is trained for $150,000$ steps with a learning rate of $3e-4$ and exponential learning rate decay.

Evaluation Details. For each scene, we evaluate the predictions from 200 frames. The predicted point maps are first projected into 3D, after which we apply farthest-point sampling to downsample the resulting point cloud for efficiency. We then aggregate the embeddings of the sampled points with those of the rejected points in their respective local neighborhoods. Finally, we compute 1-nearest-neighbor correspondences to the ground-truth point cloud to obtain the evaluation metrics.

To align the predicted point cloud with the ground-truth coordinate system, we fix the reference frame using the ground-truth camera pose of the first image, consistent with VGGT's convention of treating the first camera as the reference. We estimate the global scale factor for alignment by solving a least-squares problem over the predicted and ground-truth camera poses.

Table 5. **Runtime and peak GPU memory usage across different numbers of input frames.** Runtime is measured in seconds, and GPU memory usage is reported in gigabytes. The input resolution is 378×448.

Input Frames	2	10	20	100	200
Time	~5s	~7s	~9s	~23s	~34s
Mem.	~2GB	~4GB	~14GB	~21GB	~40GB

Table 6. **Depth fine-tuning ablation.** We evaluate depth prediction after fine-tuning VGGT [50] on ScanNet using geometric and semantic losses.

	AbsRel↓	$\delta_{1.25}$↑
VGGT (pre-trained)	0.260	63.31
VGGT (fine-tuned geometry)	0.256	63.52
VGGT (fine-tuned geometry & semantics)	**0.251**	**63.77**

B. Compute and Inference Time

As detailed in Tab. 5, we report inference runtime and peak GPU memory for the full forward pass, including the semantic, instance, and articulation heads, across varying numbers of input frames. All measurements are obtained on a single NVIDIA H100 GPU using a JAX implementation with flash attention. Input images are rendered at a resolution of 378×448. To balance inference speed and memory consumption, the DPT heads are executed sequentially for each output image, which reduces peak memory usage.

C. Geometry and Semantic Synergy

The aim of our work is to bring semantic information into models that mainly rely on geometric signals. For this purpose we are using the pre-trained geometric features of VGGT [50] as a strong starting point for learning semantics. We also expect that learning semantic features can, in turn, support geometric tasks. To test this idea, we run a small

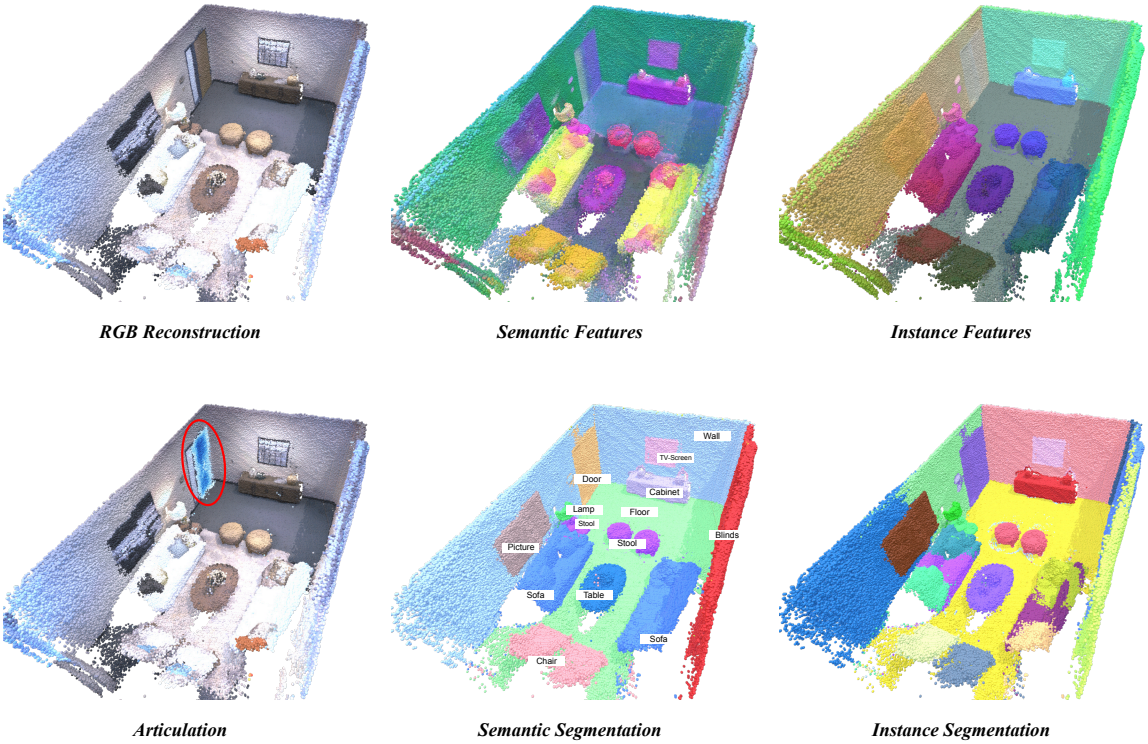

Figure 4. **Results of UNITE on Replica.** UNITE not only produces accurate geometric reconstruction, but also highly distinctive semantic and instance features. These allow for accurate semantic segmentation and instance segmentation results.

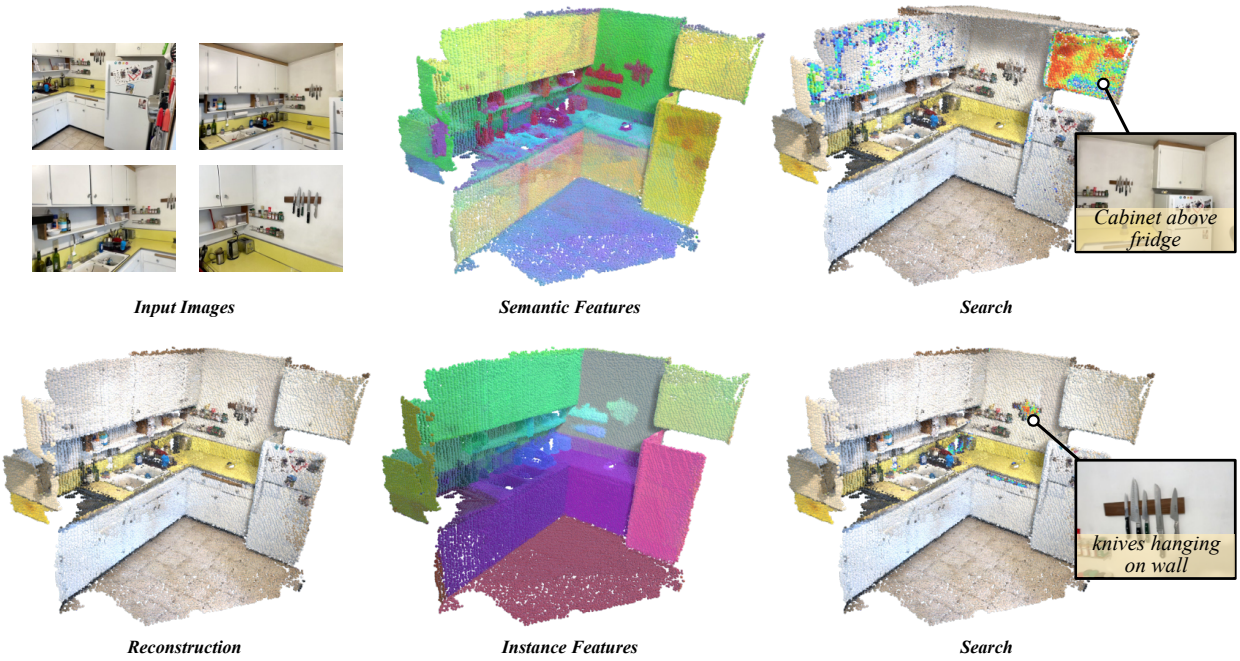

Figure 5. **Results of UNITE on *in-the-wild* scene.** UNITE is able to produce high-quality geometry, instance, and semantic features for *in-the-wild* scenes. The predicted features allow for text-based queries of high complexity, correctly identifying objects from relationship descriptions.

Table 7. **Replica Evaluation.** We evaluate UNITE on Replica for 3D semantic segmentation. The Replica dataset contains 51 semantic classes.

	mIoU	mAcc
ConceptFusion [17]	10.0	17.0
ConcpetGraphs [12]	13.0	21.0
HOV-SG [54]	14.4	21.2
OpenScene [37]	15.9	24.6
Ours	**17.0**	**26.6**

ablation on depth estimation and compare VGGT's depth predictions in three setups: the original pre-trained model, fine-tuning on ScanNet [3] with only geometric losses for $10,000$ steps, and fine-tuning on ScanNet with both geometric and semantic losses for $10,000$ steps.

As shown in Tab. 6, fine-tuning with geometric losses brings only a slight gain over the pre-trained model. This is reasonable since the model is already strong in geometric reasoning and was already trained on ScanNet. Notably, introducing our proposed semantic losses adds small improvements also in depth prediction, indicating that jointly learning geometric and semantic representations could also benefit geometric learning. However, the margins are still relatively small but we believe these results could also transfer and scale when trained on larger datasets.

D. Out-of-Distribution Evaluation

Replica. In Tab. 7, we report 3D semantic segmentation performance on the Replica dataset [44] following the same evaluation protocol as for ScanNet [3] and ScanNet++ [57]. This experiment highlights the out-of-distribution generalization capability of our method, which surpasses existing open-vocabulary baselines. In Fig. 4, we provide additional, qualitative results.

LERF Scenes. In Fig. 5, we present predictions on the kitchen scene from the *in-the-wild* LERF dataset [22]. UNITE adapts robustly to such unconstrained settings, accurately grounding complex textual queries such as *cabinet above the refrigerator* or *knives hanging on the wall*. It further distinguishes fine-grained spatial relationships, for instance, correctly differentiating the cabinets beneath the kitchen counter and those above the sink, from the cabinet above the refrigerator.