# TOMOS3D: ZERO-SHOT 3D INSTANCE SEGMENTATION FROM CUTTING PLANE RENDERINGS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We present TOMOS3D, a zero-shot method for instance segmentation in 3D scenes. Existing approaches segment input images and then match partial, noisy masks from different viewpoints, which is the central challenge dominating their design. Instead, we take a textured mesh as input and render multiple cutting plane views of it. Their number is independent of the number of input frames. Every rendered pixel maps to a 3D triangle. Each view is segmented independently and its masks lifted directly onto the mesh. The per-view SAM 2 segmentations are then stitched into one coherent global decomposition purely by mask overlap on the mesh, with no cross-view matching. To classify the objects, we propose MASKPROBE that adapts the final attention pooling of a frozen vision–language model using projected instance masks. Multiple instances thus share one image-encoder pass per image, with pooling performed separately for each instance. For evaluation, we report the fraction of ground-truth instances recovered.
TOMOS3D outperforms state-of-the-art training-free methods by a large margin, 51.5 against 43.1 on ScanNet, and comes within 2.8 points of the supervised ESAM-E on its training domain. The same setting transfers without adaptation to 3RScan and SceneNN, where TOMOS3D outperforms even the supervised approaches (with a score of 72.9 versus 59.0 for the best prior method). The MASKPROBE recognition head improves open-vocabulary recall, outperforming even MLLM-based methods, while reducing labeling time.[1]

## 1 INTRODUCTION

Class-agnostic instance segmentation of 3D scenes decomposes a reconstructed scene, typically a mesh or point cloud fused from a posed RGB-D sequence, into its constituent objects. Semantic instance segmentation additionally assigns each one a label. Machines that act in physical space, such as a robot grasping a mug, an AR headset anchoring content to furniture, or an embodied agent answering questions about a room, need to know an object's 3D extent and geometry, which part of the scene it occupies, and its semantic label.

Annotated 3D data, however, remains orders of magnitude smaller than annotated 2D data: 2D segmentors are trained on up to billions of masks (Kirillov et al., 2023; Ravi et al., 2024), and vision–language models on web-scale image–text pairs (Radford et al., 2021; Tschannen et al., 2025), while annotated 3D scans number only in the thousands. Consequently, a dominant approach to open-world 3D understanding is to transfer strong 2D models into 3D (Yang et al., 2023; Takmaz et al., 2023; Yin et al., 2024).

Existing methods segment the posed RGB(-D) frames of a scan, lift each mask into 3D through depth and camera pose, and merge hundreds of noisy partial observations into scene-level instances. This merging dominates their design: merge trees (Yang et al., 2023), view-consensus graphs (Yan et al., 2024), superpoint graphs (Yin et al., 2024), mask trackers (Zhao et al., 2025), and learned matchers (Xu et al., 2025b), the last being another component that must transfer across domains.

The proposed TOMOS3D segments the reconstructed mesh instead of the input images. The mesh is sliced with oriented planes and each side of a plane is rendered orthographically with the other side clipped away. This exposes objects from various viewpoints, including those that no recorded frame

[1] The code will be made publicly available.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




Figure 1: **TOMOS3D cuts a textured mesh with randomly oriented planes and renders each side of a plane orthographically.** Frozen SAM 2 segments the renderings, and the masks are lifted back onto the mesh to form 3D instances. *Left*: input scan and output instances, split by a cutting plane. *Right*: renderings of both sides of that plane, and a bird's-eye view.

shows (Fig. 1). Eighty such cutting-plane views summarize a scene, however many frames were recorded. A frozen SAM 2 (Ravi et al., 2024) segments each view, the renderer's per-pixel triangle identities lift every mask exactly onto the mesh, and instances are formed by clustering masks by surface overlap. SAM 2 also outputs masks at alternative granularities that prior works discard, which we keep as separate levels that roughly correspond to parts, objects, and larger regions.

To name the objects, we propose MASKPROBE, a recognition head built on the frozen vision–language model SigLIP2 (Tschannen et al., 2025). It runs on the original photographs rather than on our renderings. Mask proposals survive the rendering domain shift because they depend on shape and boundaries, which the renders preserve, but recognition depends on appearance: untextured, uniformly lit cross-sections lack the cues a vision–language model was trained on, and its labels degrade accordingly. We modify only the final pooling mechanism after the image encoder: standard masked attention (Vaswani et al., 2017) restricts pooling to image patches inside each projected 3D instance mask. All instances in a photograph therefore share a single image-encoder pass. Only the final pooling is repeated per instance, and its output is compared with text embeddings of candidate classes. By contrast, mask-conditioned encoders such as Alpha-CLIP (Sun et al., 2024) require a separate encoder pass for each instance, while OpenMask3D (Takmaz et al., 2023) runs the full encoder on multiple crops per instance. MASKPROBE reuses the pretrained pooling weights without adding parameters or training.

In class-agnostic instance proposal recall, with one hyperparameter setting for all datasets, TOMOS3D outperforms every evaluated training-free method by a large margin on ScanNet, 3RScan, and SceneNN. Supervised methods trained on ScanNet lead only on ScanNet; outside their training data, TOMOS3D matches them on 3RScan and exceeds them by a wide margin on SceneNN, while the recent ESAM collapses under the domain shift. The remaining gap to supervised methods lies at strict overlap thresholds: TOMOS3D covers nearly every annotated object, but its masks do not always follow the benchmark's exact boundaries or granularity. MASKPROBE improves open-vocabulary mean recall over OpenMask3D's crop-based recipe on all three datasets with the same SigLIP2 encoder and text prompts, while reducing the labelling time significantly.

We make the following contributions:

- TOMOS3D: a new training-free approach for 3D object instance segmentation that segments renders of a textured mesh sliced by cutting planes. TOMOS3D significantly outperforms state-of-the-art 3D instance segmentation methods.
- MASKPROBE: an efficient parameter-free recognition head that pools frozen SigLIP2 features by masked attention, requiring one encoder pass per image.

## 2 RELATED WORK

**Trained 3D instance segmentation.** Supervised segmentors such as Mask3D (Schult et al., 2023) and ISBNet (Ngo et al., 2023) learn from instance-annotated scans. Even annotation-free Seg-

ment3D (Huang et al., 2024), which trains on SAM pseudo-labels, still fits a 3D backbone to the same 1,500 ScanNet scenes (Dai et al., 2017a). In contrast, SAM is trained on 11 million images (Kirillov et al., 2023). TOMOS3D relies solely on frozen SAM2, leveraging its training scale.

**Lifting 2D masks from recorded frames.** Training-free methods apply a frozen 2D segmenter such as SAM (Kirillov et al., 2023) to the recorded RGB-D frames and lift the masks into 3D with depth and camera poses. One object thus yields many partial, occluded masks, and the methods differ mainly in how they associate them. SAM3D (Yang et al., 2023) merges back-projected masks by geometric overlap, SAI3D (Yin et al., 2024) accumulates mask evidence on a superpoint graph, MaskClustering (Yan et al., 2024) clusters masks by view consensus, SA3DIP (Yang et al., 2024a) adds pretrained 3D priors, and SAMPro3D (Xu et al., 2025a) enforces consistency earlier by projecting 3D-sampled prompts into the frames. Others delegate association to a tracker or, at the cost of training, to a learned matcher: SAM2Object (Zhao et al., 2025) and Any3DIS (Nguyen et al., 2025) propagate masks with SAM2 (Ravi et al., 2024), ESAM (Xu et al., 2025b) learns to match mask queries online, and AutoSeg3D (Wang et al., 2025) adds long- and short-term tracking memory. All of these must merge partial masks of the recorded frames, whose number grows with the scan. TOMOS3D instead segments a fixed number of renders of the mesh sliced by cutting planes, and the renderer gives exact pixel-to-triangle correspondence, so masks from different views are compared directly as triangle sets, without cross-view association. Segmenting renders has so far targeted single objects, e.g., SAMPart3D (Yang et al., 2024b), which distils 2D features of object renders into a trained 3D backbone. TOMOS3D segments plane-cut renders of whole scenes and uses the 2D masks as they are, with no 3D backbone to train.

**Open-vocabulary 3D instance recognition.** OpenScene (Peng et al., 2023) assigns a language-aligned feature to every 3D point. Instance-level methods instead separate proposal generation from recognition, following OpenMask3D (Takmaz et al., 2023): each Mask3D proposal is projected into its most visible frames and described by CLIP features aggregated over multi-scale crops. Open3DIS (Nguyen et al., 2024), Open-YOLO 3D (Boudjoghra et al., 2025), and Details Matter (Jung et al., 2025) change the proposals, the pooling, or the 2D model within this recipe. OpenTrack3D (Zhou et al., 2025) replaces similarity matching with a multimodal LLM, which must run for every proposal and text query. We follow this separation of 3D proposals and recognition with a 2D vision-language model, but replace multi-view crops with an efficient, context-aware recognition head based on masked attention pooling (Sec. 4).

**Mask-conditioned features.** Alpha-CLIP (Sun et al., 2024), used for 3D masks by Details Matter (Jung et al., 2025), fine-tunes CLIP to take the mask as an alpha channel: the weights do not carry over to newer encoders, and every image–mask pair needs its own encoder pass. MASKPROBE encodes each photograph once and restricts SigLIP2's frozen attention-pooling head to the projected mask by standard masked attention. TextRegion (Xiao et al., 2025) uses similar masking in 2D, and CAMFusion (Berriel Martins et al., 2026) on projected 3D instances, with a trained multi-view fusion. Masked attention is also expected to improve methods such as LangSplat (Qin et al., 2024), which encode each mask as a separate crop.

## 3 3D INSTANCE SEGMENTATION

TOMOS3D segments a textured triangle mesh, fused from a posed RGB-D sequence by standard reconstruction (Dai et al., 2017b), into class-agnostic instances at three granularities that roughly correspond to parts, objects, and regions. It is training-free, relying only on an off-the-shelf 2D segmenter, SAM 2 (Ravi et al., 2024), with no 3D training data, detector, or scene-specific tuning.

The idea is to run the 2D segmenter not on the recorded sequence but on the reconstructed surface itself. We slice the mesh with a set of oriented planes and render each half-space *orthographically*, so every plane yields two clean images of the mesh with the far side clipped away, and SAM 2 proposes object masks in each image. Every object appears in the renderings of every plane (on one side, or both when the cut passes through it), but each view yields only a partial, noisy fragment: foreshortened or occluded in oblique views, and split in two at a cut. The core of the method is to glue these fragments back together into coherent 3D instances. Each mask lifts back exactly, through a per-pixel triangle-identity buffer, to a set of triangles on one shared surface, so gluing needs no cross-view association: an area-weighted affinity graph over the lifted sets and a stability-based clustering with no preset number of objects suffice. Fig. 2 summarizes the pipeline. Alg. 1

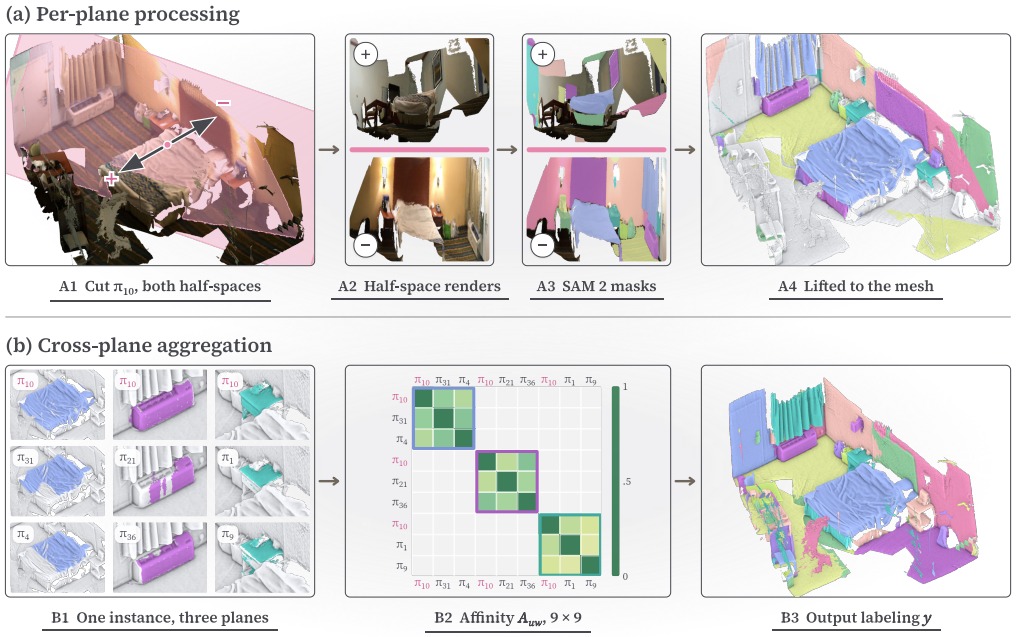



Figure 2: **TOMOS3D pipeline: segment cutting-plane views, aggregate on the mesh. (a) Per-plane processing.** Each of the $K$ sampled planes cuts the mesh into two half-spaces (A1), rendered orthographically (A2) and segmented by a frozen SAM 2 (A3). A triangle-id buffer maps each mask exactly to mesh triangles (A4). Masks on opposite sides of the plane are stitched through shared cut triangles. **(b) Cross-plane aggregation.** Components from different planes describe overlapping portions of the same instances (B1). Their area-weighted Jaccard overlap defines an affinity graph (B2), whose stable clusters yield 3D instances by triangle voting (B3). Aggregation runs independently for each of SAM 2's three output mask granularities (omitted here for simplicity), producing fine-to-coarse labelings in a single segmentation pass per view.

in the supplementary lists it step by step, and all thresholds and counts below are hyperparameters defined in Tab. 4 there.

**Input.** A triangle mesh $\mathcal{M} = (V, F)$ with vertex set $V \subset \mathbb{R}^3$ and triangle set $F$. We write $t \in F$ for a triangle, $a(t)$ for its area, and extend the latter to any subset $\mathcal{G} \subseteq F$ by $a(\mathcal{G}) = \sum_{t \in \mathcal{G}} a(t)$.

**Output.** Three labelings $y^0, y^1, y^2 : F \to \mathbb{Z}_{\geq 0}$ of the mesh, one per granularity from finest to coarsest. Each gives every triangle an instance index, with $y^\ell(t) = 0$ for unlabeled triangles. Secs. 3.1–3.3 construct a single such labeling, written $y$. Sec. 3.4 repeats the construction per granularity.

### 3.1 GEOMETRIC DECOMPOSITION

This stage converts the mesh into clean 2D images for the segmenter, each paired with an exact pixel-to-triangle map: per plane $k$ and side $s$ it outputs a mask set $\Omega_{k,s}$ and a triangle-id buffer $\Theta_{k,s}$.

**Stratified plane sampling.** We sample a fixed number $K$ of planes $\mathcal{P} = \{\pi_k\}_{k=1}^K$, so a scene is represented by $2K$ renderings, a budget independent of the length of the recorded sequence (tens of views in place of the hundreds to thousands of frames in a scan). A plane $\pi_k = (n_k, p_k)$ has a unit normal $n_k$ and a point $p_k$, and separates two half-spaces (sides) $s \in \{+, -\}$. With so few views, each surface should still appear roughly head-on and unoccluded in at least one rendering, since oblique views foreshorten objects and degrade the masks. We therefore draw the planes from two strata. The first $n_{\text{bev}}$ planes are horizontal bird's-eye views, which show floor-supported and ceiling-mounted structure head-on. The remaining $K - n_{\text{bev}}$ planes have normals drawn uniformly on the sphere and pass through the bulk of the geometry, where a cut separates objects and exposes interiors

that no exterior viewpoint shows (e.g., the floor beneath a table). Sec. A.1 gives the placement of both strata.

**Half-space rendering.** An orthographic camera placed on the cut plane sees exactly one half-space: for side $+$ it looks along $+n_k$ with the near plane on $\pi_k$ and the far plane just beyond the farthest vertex of that side, so geometry across the cut is clipped. Side $-$ is rendered identically with $n_k$ replaced by $-n_k$. Rasterizing side $s$ of $\pi_k$ yields an RGB image $I_{k,s}$ and an aligned triangle-id buffer $\Theta_{k,s} \in (\{-1\} \cup F)^{H \times W}$, where $\Theta_{k,s}(p)$ is the triangle visible at pixel $p$ ($-1$ if none). The renders are consumed as-is: we do not re-project the scan's posed RGB-D frames onto them.

**2D mask proposals.** We segment each image with SAM 2 (Ravi et al., 2024), querying its automatic mask generator on a regular prompt grid restricted to foreground pixels ($\Theta_{k,s} \geq 0$). A mask is kept if its predicted IoU, its stability score (Kirillov et al., 2023), and its share of the image area clear minimum thresholds. The retained binary masks on $I_{k,s}$ form the set $\Omega_{k,s}$.

### 3.2 Multi-view Instance Consensus

This stage relates the masks on the surface rather than in image space: each mask is lifted to a triangle set, the two sides of every cut are stitched into per-plane components, and components from different planes are compared by their overlap on the mesh, giving a sparse affinity graph that Sec. 3.3 clusters into instances.

**Mask lifting.** A mask $m \in \Omega_{k,s}$ lifts to the triangles visible through it in the triangle-id buffer,

$$\mathcal{T}(m) = \{\, t \in F : \exists\, p,\ m(p) = 1 \wedge \Theta_{k,s}(p) = t \,\}. \tag{1}$$

Triangles claimed by no mask stay unassigned and never enter the graphs below.

**Cross-cut stitching.** An object straddling a cut appears as two masks, one per side, that must be recognized as the same object. We stitch them through the *cut triangles* $\mathcal{C}_k \subseteq F$, those with vertices on both sides of $\pi_k$. They are rendered, in part, on both sides, so a cut triangle shared by a $+$ mask and a $-$ mask witnesses that they describe one object. We join a cross-side pair $(m, m') \in \Omega_{k,+} \times \Omega_{k,-}$ when the number of cut triangles they share is significant both absolutely and relatively,

$$\big|\mathcal{C}_k \cap \mathcal{T}(m) \cap \mathcal{T}(m')\big| \ \geq\ \max\big(\tau_{\text{cnt}},\ \lfloor \tau_{\text{frac}} \cdot \min(|\mathcal{T}(m)|, |\mathcal{T}(m')|) \rfloor\big); \tag{2}$$

Union–find over $\Omega_{k,+} \cup \Omega_{k,-}$ with these edges yields the per-plane components; component $u$ has triangle set $\mathcal{T}(u) = \bigcup_{m \in u} \mathcal{T}(m)$ (threshold roles in Sec. A.1).

**Cross-plane affinity.** Two components from different planes that cover nearly the same triangles are almost certainly the same object. Pooling the components of all $K$ planes into $\mathcal{U}$, we measure this with an area-weighted Jaccard affinity

$$A_{uw} = \frac{a\big(\mathcal{T}(u) \cap \mathcal{T}(w)\big)}{a\big(\mathcal{T}(u) \cup \mathcal{T}(w)\big)}, \qquad u, w \in \mathcal{U}, \tag{3}$$

which, unlike a triangle count, is invariant to tessellation density. $A$ is built sparsely through an inverted index from triangles to the components containing them, and entries below $A_{\min}$ are dropped.

### 3.3 Stable Instance Extraction

This stage turns the affinity graph into the labeling $y$. Objects live at very different scales in this graph (a compact chair forms a tight, high-affinity group, while a long sofa seen edge-on is looser) so no single global threshold separates them. We instead keep the *most stable* groups across all thresholds, then label triangles by agreement within each group.

**Stability clustering.** We convert affinity to a pseudo-distance $D_{uw} = 1 - A_{uw}$ ($D_{uw} = 1$ for pairs pruned by $A_{\min}$), build a single-linkage dendrogram over $\mathcal{U}$, and select clusters from it by HDBSCAN's Excess-of-Mass criterion (Matas et al., 2004; Campello et al., 2015). Scales are read as affinities, $\lambda = 1/D$. A node $\nu$ of the dendrogram is born at $\lambda_b(\nu)$, where its children merge, and merges into its parent at the lower scale $\lambda_d(\nu)$; read in the direction of increasing $\lambda$, $\lambda_d(\nu)$ is the scale at which $\nu$ separates from its parent, and so takes the role of the cluster birth in the

usual condensed-tree convention. Writing $\lambda_b(u)$ for the scale at which leaf $u$ separates from $\nu$, the stability of $\nu$,

$$\mathrm{stab}(\nu) = \sum_{u \in \mathrm{leaves}(\nu)} \big(\lambda_b(u) - \lambda_d(\nu)\big) \tag{4}$$

is the total affinity range over which its members persist as one group. Excess-of-Mass returns the antichain of nodes maximizing total stability: a flat clustering $\mathcal{S} = \{S_1, \ldots, S_J\}$ with no preset number of clusters. Each cluster must draw on at least $\mu$ components, which suppresses single-view oversegmentations; near-duplicate clusters separated only by affinity noise are merged with a selection tolerance $\epsilon_{\mathrm{cs}}$.

**Triangle assignment.** Each cluster $S_j$ becomes a surface region by majority vote over its components. Let $\kappa(t, j) = |\{u \in S_j : t \in \mathcal{T}(u)\}|$ count the components of $S_j$ that claim triangle $t$. We accept $t$ for $S_j$ only above a size-relative quorum $\tau_j$, and assign each triangle to its best cluster,

$$y(t) = \arg \max_{j\,:\,\kappa(t,j) \geq \tau_j} \frac{\kappa(t, j)}{|S_j|}, \qquad \tau_j = \max\big(\tau_{\min}, \lceil q\,|S_j| \rceil\big); \tag{5}$$

triangles meeting no quorum keep $y(t) = 0$.

### 3.4 Hierarchical Segmentation

SAM 2's decoder predicts three masks per prompt, following the multi-mask design introduced in SAM to handle prompt ambiguity: a single point can refer to several valid extents (Kirillov et al., 2023). These outputs roughly correspond to part-, object-, and region-level extents, without guaranteed semantic roles or strict nesting. We index them by a level $\ell \in \{0, 1, 2\}$ from finest to coarsest and retain all three, applying non-maximum suppression *within* each level so that the granularities never suppress one another; this gives three mask sets $\Omega^{\ell}_{k,s}$ from the *same* renderings. Running Secs. 3.2–3.3 once per level yields three coupled labelings $y^0, y^1, y^2$ of one mesh. The planes, renderings, and triangle-id buffers are shared and SAM 2 is invoked once per rendering, so the extra levels cost only the (inexpensive) graph construction and clustering.

**Unranked output.** TOMOS3D returns $y^0, y^1, y^2$ as *unranked* instance sets and attaches no confidence score: a training-free segmenter has no learned score head, so any ranking would be a bolted-on heuristic. Sec. 5 gives the matching evaluation protocol.

## 4 Semantic Labelling

We assign semantic labels to the class-agnostic instances of Sec. 3 with a frozen image–text encoder, without modifying the masks or any additional training. Note that instance discovery operates on synthetic cross-section renderings, whereas classification uses the original posed RGB frames. The two tasks tolerate the rendering domain shift very differently. Mask proposals rely on shape and boundaries, which the renders reproduce faithfully, so a promptable segmenter transfers to them. Labelling relies on appearance, and the renders' flat shading and missing photometric context move them away from the images an image–text encoder was trained on. We therefore keep the recognition head on the photographs.

Let $\mathbf{m} \subseteq F$ denote an instance. Following OpenMask3D (Takmaz et al., 2023), we select the $n_{\mathrm{view}}$ RGB frames in which $\mathbf{m}$ is most visible, measured by the number of its mesh vertices that project into the frame unoccluded, and discard frames with fewer than $n_{\min}$ such vertices. This selection is shared by all heads we compare. Projecting the vertices of $\mathbf{m}$ into a frame yields a sparse point set rather than a region, so we rasterize it into a binary mask $M_v$, $v \in \{1, \ldots, n_{\mathrm{view}}\}$: we splat the $N_v$ visible vertices, apply a morphological closing with an elliptical kernel of radius

$$r_v = \mathrm{clip}\Big(\big\lceil \alpha \sqrt{A_v / N_v}\,\big\rceil,\, r_{\min},\, r_{\max}\Big) \tag{6}$$

pixels, where $A_v$ is the area of the points' bounding box, and dilate the result with a fixed $d \times d$ kernel. The radius follows the mean spacing of the projected vertices, so close-up views are closed with small kernels and distant, sparsely sampled ones with large kernels. Each class $c$ has a unit-normalized text prototype averaged over a prompt ensemble $\mathcal{T}$ of templates (*e.g.* "This is a photo of

Table 1: **Class-agnostic recall at 500-proposal budget.** Mask3D (top row) uses its default setting; Mask3D (tuned) keeps its untruncated candidate set. Supervised methods are trained on ScanNet; their 3RScan and SceneNN results measure zero-shot transfer. Recall metrics at looser overlap thresholds are also shown. **Best** and second best training-free method highlighted.

| | Method | SceneNN | | | 3RScan | | | ScanNet v2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AR@500 | $AR_{50}$ | $AR_{25}$ | AR@500 | $AR_{50}$ | $AR_{25}$ | AR@500 | $AR_{50}$ | $AR_{25}$ |
| *3D supervised (ScanNet)* | Mask3D (Schult et al., 2023) | 56.5 | 77.2 | 90.2 | 36.5 | 52.5 | 71.2 | 62.8 | 79.2 | 91.1 |
| | Mask3D (tuned) | 59.0 | 79.7 | 90.2 | 44.3 | 63.9 | 82.1 | 70.0 | 85.6 | 94.5 |
| | ESAM | 50.6 | 75.6 | 90.2 | 13.9 | 27.9 | 58.8 | 53.4 | 72.8 | 87.8 |
| | ESAM-E | 52.8 | 75.6 | 92.7 | 13.9 | 26.5 | 62.3 | 54.3 | 73.4 | 88.1 |
| *Training-free* | SAM3D | 29.1 | 56.9 | 89.4 | 8.2 | 22.1 | 56.6 | 19.1 | 42.2 | 70.4 |
| | SAI3D | 19.7 | 43.1 | 88.6 | 21.6 | 39.4 | 73.5 | 43.1 | 65.0 | 87.8 |
| | SAM2Object | 37.8 | 61.0 | 82.1 | 16.3 | 27.2 | 49.6 | 37.7 | 56.3 | 76.9 |
| | TOMOS3D (ours) | **72.9** | **91.9** | **98.4** | **41.4** | **65.9** | **91.8** | **51.5** | **74.5** | **94.4** |

a {}."), $\mathbf{t}_c = \mathcal{N}\big(\frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}}\mathcal{N}(\phi_{\text{txt}}(\tau(c)))\big)$, where $\mathcal{N}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$, and an instance embedding $\mathbf{z}$ is classified by cosine similarity, $\hat{c} = \arg\max_c \langle \mathbf{z}, \mathbf{t}_c \rangle$. What remains is the construction of $\mathbf{z}$.

**Masked attention pooling.** Rather than averaging embeddings of multi-scale crops, which discards the mask geometry, we restrict the encoder's own pooling to the patches covered by the projected mask, keeping the encoder frozen. SigLIP2 (Zhai et al., 2023; Tschannen et al., 2025) aggregates patch tokens $X \in \mathbb{R}^{N\times D}$ with a multi-head attention pooling (MAP) head in which a learned probe $\mathbf{p}$ attends to the tokens: with $\mathbf{q} = W_q\mathbf{p} + \mathbf{b}_q$, $K = W_kX + \mathbf{b}_k$, $V = W_vX + \mathbf{b}_v$, and logits $\boldsymbol{\ell} = \mathbf{q}K^\top/\sqrt{d}$, the image embedding is $\phi_{\text{img}}(I) = \text{MLP}(\text{LN}(W_o\,\text{softmax}(\boldsymbol{\ell})V))$. We area-pool $M_v$ onto the patch grid to obtain per-patch coverage $c_i \in [0,1]$, take $\mathcal{K}_v = \{i : c_i \geq \theta\}$ (or $\arg\max_i c_i$ if empty), and mask the logits before the pooling softmax:

$$\mathbf{z}_v = \mathcal{N}\Big(\text{MLP}\big(\text{LN}(W_o\,\text{softmax}(\tilde{\boldsymbol{\ell}})V)\big)\Big), \qquad \tilde{\ell}_i = \begin{cases} \ell_i, & i \in \mathcal{K}_v, \\ -\infty, & \text{otherwise.} \end{cases} \tag{7}$$

The instance embedding $\mathbf{z}$ is the normalized mean of the $n_{\text{view}}$ $\mathbf{z}_v$. Setting $\mathcal{K}_v = \{1,\ldots,N\}$ exactly recovers the encoder's standard pooling. We call this head MASKPROBE, since the projected mask is what restricts the encoder's own pooling probe. MASKPROBE-crop applies the same masked pooling to a crop rather than to the full frame.

## 5 EXPERIMENTS

**Datasets.** We evaluate class-agnostic 3D instance segmentation on three RGB-D scan datasets: the 312-scene validation split of ScanNet v2 (Dai et al., 2017a), the 46-scene validation split of 3RScan (Wald et al., 2019), and the 12-scene transfer split of SceneNN (Hua et al., 2016) with the 18 NYU "thing" classes used by ESAM (Xu et al., 2025b). For open-vocabulary instance segmentation, we use the ScanNet200 vocabulary (Rozenberszki et al., 2022), the RIO27 thing classes on 3RScan, and the same 18 NYU classes on SceneNN. Sec. A gives the mesh and frame details.

**Primary metric.** We evaluate proposal coverage using average recall with at most 500 proposals per scene (AR@500). Recall is the fraction of ground-truth instances matched by a proposal at a given IoU threshold; AR@500 averages recall over thresholds from 0.50 to 0.90 in steps of 0.05. $AR_{50}$ and $AR_{25}$ report recall at IoU thresholds of 0.50 and 0.25, respectively. For semantic evaluation, matches must also have the correct class; we additionally average recall over classes and denote the result mAR@500. This choice reflects the ambiguity of annotation granularity: a zero-shot method cannot assume a benchmark's preferred level, so TOMOS3D returns an unranked union of three granularities. Recall measures coverage without penalizing valid predictions at unannotated granularities or outside the annotated object set, while the proposal budget constrains over-generation without requiring calibrated confidence scores. AR@$K$ is standard for 2D object proposals (Hosang et al., 2016; Lin et al., 2014) and is also used by SAM (AR@1000 on LVIS) to evaluate open-world masks beyond the annotated classes (Kirillov et al., 2023). Sec. B gives the formal definition.

**Baselines.** We compare TOMOS3D with the ScanNet-supervised Mask3D (Schult et al., 2023) (reference top-200 proposal variant, and a tuned variant retaining all proposals, which the text dis-

Table 2: **Open-vocabulary recognition on TOMOS3D proposals.** All rows label the same class-agnostic proposals. OM3D is OpenMask3D's original head (Takmaz et al., 2023); for OM3D$^+$, we update OM3D so that it shares MASKPROBE's SigLIP2 encoder and prompt ensemble. Acc is micro accuracy (%) on matched instances (Sec. B). Best per column in **bold**, second underlined.

| | SceneNN | | 3RScan | | ScanNet200 | | Avg |
|---|---|---|---|---|---|---|---|
| Readout | $mAR_{500}$ ↑ | Acc ↑ | $mAR_{500}$ ↑ | Acc ↑ | $mAR_{500}$ ↑ | Acc ↑ | $mAR_{500}$ ↑ |
| OM3D | 58.2 | 78.8 | 19.2 | 46.4 | 23.7 | 26.1 | 33.7 |
| OM3D$^+$ | 60.7 | 78.8 | 24.4 | **59.8** | 34.0 | 48.8 | 39.7 |
| Qwen3-VL-4B | 62.2 | 82.3 | 22.6 | 58.3 | 33.2 | **58.6** | 39.3 |
| MASKPROBE (ours) | **63.1** | **83.2** | **25.8** | **59.8** | **35.4** | 52.3 | **41.4** |

cusses) and ESAM/ESAM-E (Xu et al., 2025b), and the training-free SAM3D (Yang et al., 2023), SAI3D (Yin et al., 2024), and SAM2Object (Zhao et al., 2025). Every class-agnostic baseline is run with its official codebase and published inference settings. The MASKPROBE semantic head is compared with OpenMask3D's multi-crop head (Takmaz et al., 2023) in two forms: OM3D, the original recipe (CLIP ViT-L/14-336, prompt "a {} in a scene"), and OM3D$^+$, our improved variant with the same SigLIP2 encoder and prompt ensemble as MASKPROBE, so that the two differ only in the readout. We also tuned a generative Qwen3-VL-4B baseline, which picks one class per view from a crop with the instance marked by an ellipse, followed by a weighted vote over views (Sec. A.2). All baseline numbers are our reproductions.

**TOMOS3D and MASKPROBE implementation.** TOMOS3D uses the frozen SAM 2.1 Hiera-Large checkpoint (Ravi et al., 2024) for masks and MASKPROBE the frozen SigLIP2-so400m (Tschannen et al., 2025); neither is trained on 3D data. Tab. 4 (Sec. A) lists all hyperparameters. Sec. C.2 ablates encoder, pooling, and prompts, and compares heads across mask sources. Every design choice is made on ScanNet and then fixed: the segmentation hyperparameters on ten held-in scenes, and the encoder, pooling, prompt ensemble, and view selection on ScanNet200, for MASKPROBE and for every head we compare it with. The sole exceptions are three dense-mesh settings on 3RScan and SceneNN; no result on those splits, in the main paper or the supplementary, informs any choice.

### 5.1 RESULTS

**Class-agnostic instance segmentation.** TOMOS3D, with one configuration fixed on ScanNet, is the best training-free method in every cell of Tab. 1 and, outside its tuning domain, beats every ScanNet-supervised method on five of six metrics. On ScanNet, TOMOS3D reaches 51.5 AR@500 and 74.5 $AR_{50}$, 8.4 AR@500 above the strongest training-free baseline (SAI3D) and 2.8 below ESAM-E. TOMOS3D covers the same objects as the best supervised model (94.4 versus 94.5 $AR_{25}$ for Mask3D (tuned)) without training on ScanNet; the remaining gap is in boundary and granularity agreement at strict overlap.

**Transfer to unseen datasets.** On 3RScan and SceneNN every method is zero-shot, the supervised ones included. TOMOS3D's AR@500 margin over the strongest training-free baseline grows to 19.8 on 3RScan (41.4 versus 21.6, SAI3D) and 35.1 on SceneNN (72.9 versus 37.8, SAM2Object); the runner-up changes with dataset and metric. Averaged over the two unseen datasets, TOMOS3D scores 57.2 AR@500 against 51.7 of Mask3D and 33.4 for ESAM-E. From ScanNet to 3RScan, ESAM-E loses 40.4 AR@500 and Mask3D (tuned) 25.7, whereas TOMOS3D loses 10.1. On 3RScan, Mask3D keeps a small 2.9-point lead in AR@500, but TOMOS3D leads at 0.5 IoU (+2.0) and 0.25 (+9.7): again, the gap closes as the threshold loosens. On SceneNN, TOMOS3D exceeds every method on every metric, by at least 13.9 AR@500; broadening the annotations beyond 18 classes widens the lead to 24.9 (Sec. C.1) . Fig. 3 shows qualitative examples; Sec. C.1 extends it.

**Open-vocabulary instance segmentation.** Tab. 2 compares recognition heads on fixed TOMOS3D proposals and Tab. 3 tests the two SigLIP2 heads across mask sources; the complete controlled matrix is in Sec. C.2. MASKPROBE has the highest mAR@500 on all three datasets (35.4, 25.8, and 63.1; 41.4 on average against 39.7 for OM3D$^+$ and 39.3 for Qwen3-VL-4B), with encoder, prompts, and view selection fixed on ScanNet200 and carried unchanged to 3RScan and SceneNN; to our knowledge, these are the first open-vocabulary instance results on 3RScan. The generative Qwen3-VL-4B trails in recall on every dataset, although it has higher micro accuracy on matched

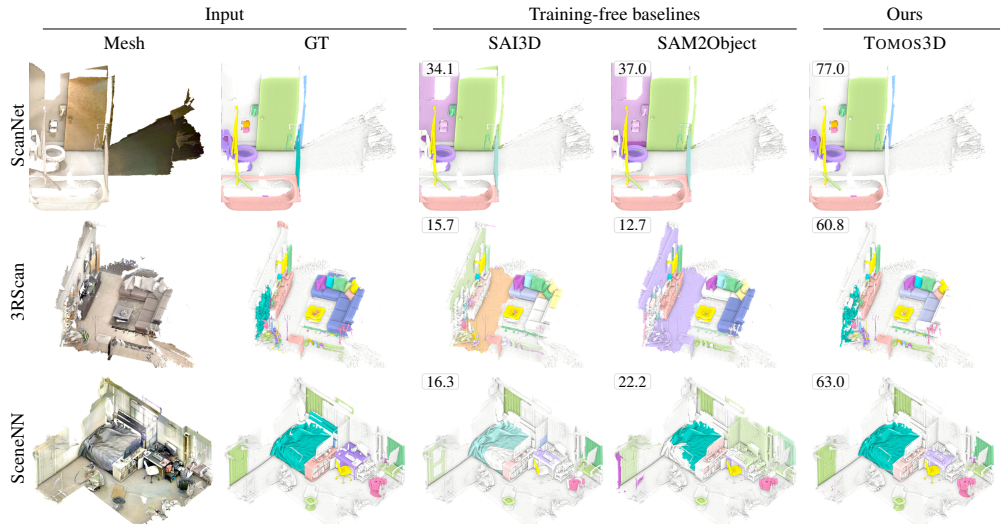



Figure 3: **Class-agnostic instance segmentation – qualitative results.** We compare with the two strongest training-free baselines of Tab. 1. Since proposals for multigranularity methods overlap, every panel shows, for each GT instance, the single proposal that matches it best; all other proposals are omitted. Each proposal takes the colour of its GT instance, lighter when its IoU is below 0.5. Numbers give AR@500 on that scene. More scenes, baselines and per-level output are in Sec. C.1.

ScanNet200 instances (58.6 versus 52.3) (Sec. C.2). The original OM3D recipe (CLIP, "a {} in a scene") reaches only 23.7 on ScanNet200 (Sec. C.2).

Mask quality sets the headroom: with MASKPROBE fixed, ground-truth masks raise mAR@500 to 66.1 / 68.3 / 80.5 and widen the gain over OM3D$^+$ from 1.4 to 5.4 points on ScanNet200 and from 1.4 to 2.1 on 3RScan (on SceneNN, the 2.4-point gain on TOMOS3D masks closes to a tie on ground truth). Masked pooling reads features inside the mask, so better proposals translate directly into recognition, while crops are less sensitive (Sec. C.2).

Table 3: **Better masks, larger MASKPROBE gain.** Cells give MASKPROBE / OM3D$^+$ on TOMOS3D and GT masks.

| Masks | ScanNet200 | 3RScan | SceneNN |
|---|---|---|---|
| TOMOS3D | 35.4 / 34.0 | 25.8 / 24.4 | 63.1 / 60.7 |
| GT masks | 66.1 / 60.7 | 68.3 / 66.2 | 80.5 / 80.5 |

**Ablations and robustness.** The supplementary ablates the hierarchy levels (Sec. C.1), validates HDBSCAN (Sec. C.1); the plane budget and sampling seed (Sec. C.1), and the recognition prompt, encoder, pooling, and mask source (Sec. C.2). Sec. C.1 analyses the proposal budget. Sec. D reports runtime and measured labelling speed-ups, and Sec. E discusses limitations in detail.

## 6 CONCLUSION

TOMOS3D demonstrates that segmenting synthetic views with known correspondence improves 3D instance segmentation. This shifts the design effort from matching observations to choosing views that expose useful scene structure. With a single setting across datasets, TOMOS3D's lead over training-free methods that segment images grows substantially beyond the training domain, suggesting that renders provide a more transferable input representation. MASKPROBE further shows that shared image features across instances can support accurate and efficient instance recognition.

**Limitations.** Broad object coverage does not yet imply precise segmentation. On ScanNet, near-parity with supervised Mask3D at loose overlap contrasts with a substantial gap at stricter thresholds, suggesting that boundary accuracy and granularity remain key limitations. A natural next step is to refine these proposals while preserving their coverage and ability to transfer across datasets.

### AI USE STATEMENT

In this work, we used generative AI coding assistants (Claude Code and Codex) for research execution: implementing parts of the method and of the experimental pipeline. We have not used generative AI tools for research ideation, for generating synthetic datasets, or for formulating mathematical claims or writing proofs; the remaining tasks with required disclosure (translation, data collection and cleaning, qualitative analysis) are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: supplementary literature retrieval and summarisation to support the authors' own literature review, drafting and revising parts of the paper text, and polishing the writing. We have reviewed all AI-assisted work and take full responsibility for the final content.

### REPRODUCIBILITY STATEMENT

TOMOS3D has no trained parameters: it uses publicly released frozen checkpoints (SAM 2.1 for segmentation, SigLIP2 for recognition) and is deterministic given the plane-sampling seed. Secs. 3 and 4 describe the method, and Alg. 1 summarises the segmentation pipeline. Tab. 4 lists the single hyperparameter setting used for all datasets, including the dense-mesh deviations on 3RScan and SceneNN. Sec. A gives the remaining implementation details: input meshes, plane sampling, rendering, cross-cut stitching, and the recognition heads and baselines. All datasets are public; Sec. B specifies the evaluation protocol, including ground-truth filtering and metric definitions. Sec. C.1 reports the variation across plane-sampling seeds, and Sec. D states the hardware and runtime. We will release the code upon publication.

## REFERENCES

Tomas Berriel Martins, Martin R. Oswald, and Javier Civera. Cross-attentive multiview fusion of vision-language embeddings. *arXiv:2604.12551*, 2026.

Mohamed El Amine Boudjoghra, Angela Dai, Jean Lahoud, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Open-YOLO 3D: Towards fast and accurate open-vocabulary 3D instance segmentation. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.

Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data*, 10(1), 2015.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017a.

Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration. *ACM Trans. Graph.*, 36(3), 2017b.

Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. What makes for effective detection proposals? *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(4):814–830, 2016.

Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. SceneNN: A scene meshes dataset with annotations. In *Proc. Int. Conf. 3D Vis. (3DV)*, 2016.

Rui Huang, Songyou Peng, Ayça Takmaz, Federico Tombari, Marc Pollefeys, Shiji Song, Gao Huang, and Francis Engelmann. Segment3D: Learning fine-grained class-agnostic 3D segmentation without manual labels. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

Sanghun Jung, Jingjing Zheng, Ke Zhang, Nan Qiao, Albert Y. C. Chen, Lu Xia, Chi Liu, Yuyin Sun, Xiao Zeng, Hsiang-Wei Huang, Byron Boots, Min Sun, and Cheng-Hao Kuo. Details matter for indoor open-vocabulary 3D instance segmentation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023.

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Trans. Graph.*, 39(6), 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014.

Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11), 2017.

Tuan Duc Ngo, Binh-Son Hua, and Khoi Nguyen. ISBNet: A 3D point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023.

Phuc 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

Phuc Nguyen, Minh Luu, Anh Tran, Cuong Pham, and Khoi Nguyen. Any3DIS: Class-agnostic 3D instance segmentation by 2D mask tracking. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

Songyou Peng, Kyle Genova, Chiyu Max Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D language gaussian splatting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv:2408.00714*, 2024.

Dávid Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3D semantic segmentation in the wild. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask transformer for 3D semantic instance segmentation. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-CLIP: A CLIP model focusing on wherever you want. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-vocabulary 3D instance segmentation. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. RIO: 3D object instance re-localization in changing indoor environments. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019.

Hanshi Wang, Zijian Cai, Jin Gao, Yiwei Zhang, Weiming Hu, Ke Wang, and Zhipeng Zhang. Online segment any 3D thing as instance tracking. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2025.

Yao Xiao, Qiqian Fu, Heyi Tao, Yuqun Wu, Zhen Zhu, and Derek Hoiem. TextRegion: Text-aligned region tokens from frozen image-text models. *Trans. Mach. Learn. Res. (TMLR)*, 2025.

Mutian Xu, Xingyilang Yin, Lingteng Qiu, Yang Liu, Xin Tong, and Xiaoguang Han. SAMPro3D: Locating SAM prompts in 3D for zero-shot instance segmentation. In *Proc. Int. Conf. 3D Vis. (3DV)*, 2025a.

Xiuwei Xu, Huangxing Chen, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. EmbodiedSAM: Online segment any 3D thing in real time. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025b.

Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. MaskClustering: View consensus based mask graph clustering for open-vocabulary 3D instance segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

Xi Yang, Xu Gu, Xingyilang Yin, and Xinbo Gao. SA3DIP: Segment any 3D instance with potential 3D priors. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024a.

Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. SAM3D: Segment anything in 3D scenes. *arXiv:2306.03908*, 2023.

Yunhan Yang, Yukun Huang, Yuan-Chen Guo, Liangjun Lu, Xiaoyang Wu, Edmund Y. Lam, Yan-Pei Cao, and Xihui Liu. SAMPart3D: Segment any part in 3D objects. *arXiv:2411.07184*, 2024b.

Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. SAI3D: Segment any instance in 3D scenes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023.

Jihuai Zhao, Junbao Zhuo, Jiansheng Chen, and Huimin Ma. SAM2Object: Consolidating view consistency via SAM2 for zero-shot 3D instance segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

Zhishan Zhou, Siyuan Wei, Zengran Wang, Chunjie Wang, Xiaosheng Yan, and Xiao Liu. OpenTrack3D: Towards accurate and generalizable open-vocabulary 3D instance segmentation. *arXiv:2512.03532*, 2025.