# CROSS-VIEW INSTANCE ASSEMBLY IN NATURAL-COLOR TOMOGRAPHY

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

## ABSTRACT

Existing segment-everything methods were built for video, three-dimensional scenes, grayscale CT data, and artificially colored microscopy. Natural-color tomography, which reveals the interior of opaque materials in RGB, is not among these domains. How well those methods transfer to this domain is largely unknown. We propose WINZE, a class-agnostic instance segmentation method designed for natural-color volumes. WINZE reads the volume section by section in three orthogonal orientations with a frozen SAM 2 backbone and assembles the sections into volumetric instances by rule without training. We compared it with seven segment-everything methods on 1,239 manually annotated instances. WINZE ranks first in instance F1@0.5 in four of the six scored scopes and achieves scores 1.3 to 3.1 times those of the best compared method on the three sets of scored planes. Our results give a reference point for this domain, motivate methods that build the spatial continuity of objects into their assumptions, and suggest that the main difficulty lies in holding each object as one label across sections.

## 1 INTRODUCTION

Natural-color tomographic volumes show the hidden interior as an RGB space (Figure 1; Iba et al., 2025). The objects inside, such as fossils embedded in rock, have to be separated from the surrounding material before anything can be measured, described, or compared. Manual segmentation does not scale. Dense annotation of three-dimensional volumes is prohibitively time-consuming (Zhou et al., 2025). Segmentation is still a bottleneck for researchers (Archit et al., 2025). Fully automatic segmentation is the key to reading what these volumes contain.

Foundation models for segmentation have advanced quickly since the SAM series (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026). There are also many methods targeting specific domains. Some are designed for grayscale medical or material CT (He et al., 2025; Zhang et al., 2026), some for stained or fluorescent microscopy (Archit et al., 2025; Pachitariu et al., 2025), and



DS00007_cube0001 DS00068_cube0002 DS00136_cube0004

Figure 1: Three of the natural-color cubes used here, each of $256^3$ voxels and rendered at equal size. DS00007_cube0001 is a packstone and the other two are carbonate concretions.



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.

Table 1: The target domains that existing segment-everything methods were built for.

| Target | Form | Color | Representative segment-everything methods |
|---|---|---|---|
| Medical, material CT | Volume | grayscale | VISTA3D (He et al., 2025), SAMamba3D (Zhang et al., 2026) |
| Light and fluorescence microscopy | Volume | artificial color | Cellpose-SAM (Pachitariu et al., 2025), μSAM (Archit et al., 2025) |
| 3D scenes | Surface, point cloud | natural color | Open3DIS (Nguyen et al., 2024), MaskClustering (Yan et al., 2024) |
| Video | Video | natural color | DEVA (Cheng et al., 2023), SAM 3 (Carion et al., 2026) |
| **Natural-color tomography** | Volume | natural color | **None (this work)** |

some for surfaces in three-dimensional scenes (Nguyen et al., 2024; Yan et al., 2024). Natural-color interior volumes therefore remain almost unserved by segment-everything methods (Table 1). Appendix A reviews related work.

We propose WINZE (*Whole-volume Instance segmentation for Natural-color tomography, a Zero-shot mining Engine*) as the first segment-everything method designed for natural-color tomography.

Our contributions are summarized as follows.

- We evaluate seven existing segment-everything methods on natural-color volumes against manual annotation.
- We propose WINZE, which reads volumes with a frozen SAM 2 backbone in three orthogonal orientations and assembles the sections into volumetric instances, without training.
- Our method scores highest in most scopes against seven compared methods on 1,239 manually annotated instances.
- We show a reference point for the segmentation task on natural-color tomography and suggest that its main difficulty lies in holding each object as one label across sections.

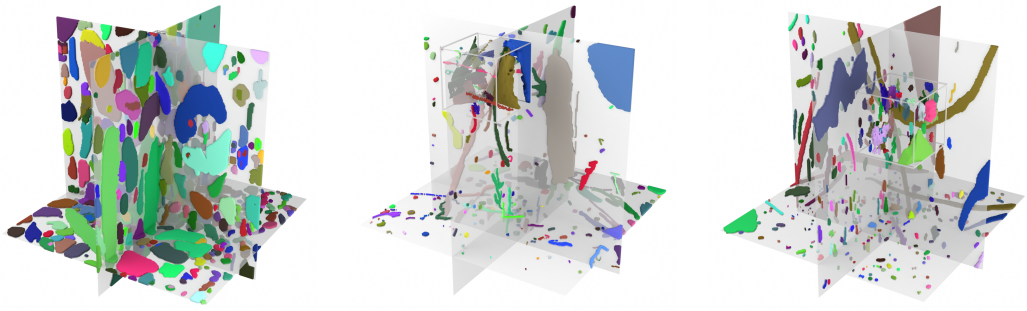
## 2 DATA

### 2.1 NATURAL-COLOR TOMOGRAPHY

The natural-color volumes were acquired with a grinding tomography system. At each step, the specimen is ground and photographed, and the volume is obtained as a sequential RGB stack (Figure 1). We used four $256^3$ cubes cropped from four volumes, `DS00007_cube0001`, `DS00068_cube0002`, `DS00136_cube0004`, and `DS00173_cube0006`. The in-plane images were downscaled to $256 \times 256$ px by a factor of 0.25, 0.2, 0.125 and 0.1667, respectively. Additionally, we used eleven public natural-color volumes for a broader comparison (Table 5; Ikegami et al., 2025).

### 2.2 GROUND TRUTH

Manual segmentation was conducted on four cubes to define ground truth using 3D Slicer 5.12.4 (Fedorov et al., 2012). Three cubes, `DS00007_cube0001`, `DS00068_cube0002`, and `DS00136_cube0004`, were annotated exhaustively within their scored planes and ROIs and used for quantitative evaluation (Figure 2). Each was annotated on three orthogonal planes, and in one ROI of $64^3$ voxels. The configuration was developed on `DS00007_cube0001` (the development cube). The two independent cubes were drawn at random, and the planes and ROIs were placed in typical regions by a fixed rule. These choices used only the raw images, and no method output entered them (Appendix B.1). The fourth, `DS00173_cube0006`, was annotated with semantic identities (e.g., bivalve shells, plants) and used for qualitative evaluation (Figure 7). Across the three cubes, these give 1,045 2D instances and 130 3D objects. With the 53 objects of the qualitative



DS00007_cube0001 DS00068_cube0002 DS00136_cube0004

Figure 2: Three manually annotated planes and one $64^3$ ROI in each of the three cubes. The ROI cube for DS00007_cube0001 is behind the planes in this view.

cube, 1,228 instances were annotated. The eleven public volumes each carry a manual label for one reference object (Ikegami et al., 2025), and together with these the comparison rests on 1,239 manually annotated instances.

## 3 From sectional evidence to volumetric instances

### 3.1 Task and assumptions

WINZE detects 2D masks with a frozen SAM 2 on $XY$, $XZ$ and $YZ$ sections, fuses their soft evidence into a volumetric foreground, links $XY$ mask pieces (atoms) along $Z$ into tracks, and groups tracklets rebuilt from these atoms into 3D objects with whole 2D masks from the three views as hyperedges. No stage is trained, and no target class or prompt is given. Two components come from u-Segment3D (Zhou et al., 2025). The consensus that turns the three per-view labelings into a provisional volumetric labeling re-implements from the published description its indirect method, which follows the summed distance-transform gradients of the three views, and the label diffusion that refines this labeling follows its per-object loop with a different update rule. The SAM 2 detection passes, the hysteresis foreground gate, the gap filling across views, the assignment and tracking of 2D atoms, the track filtering and the whole-mask grouping are specific to WINZE.

Let $I : \Omega \to \mathbb{R}^3$ be an RGB volume on the discrete lattice $\Omega = \{1, \dots, n_x\} \times \{1, \dots, n_y\} \times \{1, \dots, n_z\}$. The desired output is a label map $\widehat{Y} : \Omega \to \{0, 1, \dots, K\}$, with zero denoting background and positive integers identifying objects. Labels are arbitrary instance identifiers, not taxonomic classes. The object count is not supplied by a user.

For view $v \in \mathcal{V} = \{XY, XZ, YZ\}$, section $t$, and scale label $s \in \mathcal{S}$, write $R_{v,t,s}(I)$ for the corresponding image. The scale-label set is $\mathcal{S} = \{\text{native}, \text{mid}, \text{coarse}\}$. Of the nine pairs in $\mathcal{V} \times \mathcal{S}$, the executed schedule uses five, carried out in four detection passes because the orthogonal pair is produced by a single pass. Appendix C.1.4 gives the field of view, resolution, points per side and detection thresholds of each pass.

No stage of the chain reads the physical voxel spacing. The preparation script resamples an input to isotropic voxels only when a spacing is supplied and its anisotropy exceeds 2 to 1. No spacing was supplied for the natural-color volumes here, and none of them was resampled beyond the in-plane downscaling of Section 2. Neighborhoods, smoothing kernels and closings are all defined in voxels. The same radius therefore spans different physical distances along different axes. The four cubes used have in-plane spacings of 27.1–52.6 µm against 10–50 µm along $Z$ (Appendix B).

### 3.2 Frozen proposal generation

We express the proposal interface as

$$(\mathcal{M}_{v,t,s}, E_{v,t,s}) = F_\theta(R_{v,t,s}(I); \eta), \qquad \theta \text{ fixed}, \tag{1}$$

where $\mathcal{M}_{v,t,s}$ is a collection of candidate masks, $E_{v,t,s}$ is the retained soft evidence, and $\eta$ collects inference settings. The checkpoint is `sam2.1_hiera_tiny` with configuration `sam2.1_hiera_t.yaml`.

Soft evidence is retained rather than immediately interpreted as final object identity. A sectional mask may describe a part, a whole object, or several touching objects. Another view may help distinguish these interpretations. The evidence values are not assumed to be calibrated probabilities. They are maxima of sigmoid outputs in $[0, 1]$. No cross-view normalization is applied beyond the fusion of Appendix C.2, and no calibration analysis is reported. Because all views derive from the same acquisition and backbone, independence of view errors would be an unjustified assumption.

### 3.3 Reconciliation and assignment

Let $\mathcal{E}$ and $\mathcal{M}$ denote the collected evidence and masks. Figure 3 gives the execution order. The frozen driver, the main script of WINZE, runs the detection passes and then two rule-based CPU phases. The first combines three-view evidence into a provisional volumetric labeling through

$$\texttt{gate} \longrightarrow \texttt{xview} \longrightarrow \texttt{smooth} \longrightarrow \texttt{postxview}. \tag{2}$$

The second assigns SAM 2 masks and cleans the resulting instances through

$$\texttt{augment} \longrightarrow \texttt{atomtrack} \longrightarrow \texttt{filter} \longrightarrow \texttt{ortho} \longrightarrow \texttt{parent}. \tag{3}$$

The result of this phase is the instance map $\widehat{Y}_0$. An atom is an $XY$ mask piece kept by `augment`, and a track is a chain of atoms along $Z$. Appendices C.1 and C.2 give every rule and threshold. The `gate` stage builds the foreground of the provisional labeling $U$ by hysteresis rather than by a single threshold, and weak evidence connected to a confident seed is kept. Within the driver, $U$ decides only which label each atom votes for. The output is made of the atoms themselves, which later stages never reshape, although `filter` can remove whole tracks and `ortho` can restore some of them. The `augment` stage merges atoms from the three scales under the priority native $>$ mid $>$ coarse. The scales complement rather than compete. Large fragments of the surrounding rock (matrix) are kept as atoms. In `atomtrack`, an atom takes the label it overlaps best when that overlap is large enough, pure enough, and clearly ahead of the runner-up. Failing that, it inherits the track of the atom it overlaps at $z - 1$, and failing that again, it starts a track of its own.

### 3.4 Whole-mask object grouping

The frozen configuration applies one further stage to $\widehat{Y}_0$. It first rebuilds the fragments. All `augment` atoms are linked again along $Z$ into tracklets, using their overlap, the provisional labeling $U$ as a weak prior and the mean color of each atom, and the tracklets are then cut to the accepted foreground of $\widehat{Y}_0$. Only the foreground of $\widehat{Y}_0$ is kept, not its track identifiers. The tracklets are then taken as immutable nodes and whole two-dimensional masks in $XY$, $XZ$, and $YZ$ as hyperedges, and they are grouped into objects. A grouping is accepted when the views agree, when the fragments are observed and projected often enough to be compared, when their RGB histograms match, and when at least two views support the same decision. A pair is vetoed when two views clearly separate it or when the colors differ, and a veto also blocks every later merge that would join the two groups. Accepted pairs are merged in decreasing order of strength. Appendix C.1 gives the thresholds and the merge order. The rule accepts no ground truth, object name, manual seed, selected region, or dataset-specific threshold, and the accepted foreground is unchanged by the grouping.

The `ortho` stage of Equation 3 is a different computation from this grouping.

### 3.5 What is and is not frozen

WINZE does not update the SAM 2 backbone. Rescaling, proposal rejection, evidence aggregation, and cleanup can all encode domain-specific choices. Likewise, RGB is an input representation, not evidence that an explicit color-affinity rule improves fusion. In the executed configuration, detection receives RGB, `smooth` and the contrast score of `filter` read the mean of the three color channels, the tracklets of Section 3.4 compare the mean RGB of each atom, and the grouping reads per-fragment RGB histograms directly. Section 5.3 tests grayscale input.

We make no guarantee about connectivity, preservation of thin walls, or invariance to orientation. Such properties require either a proof under explicit assumptions or targeted measurements.

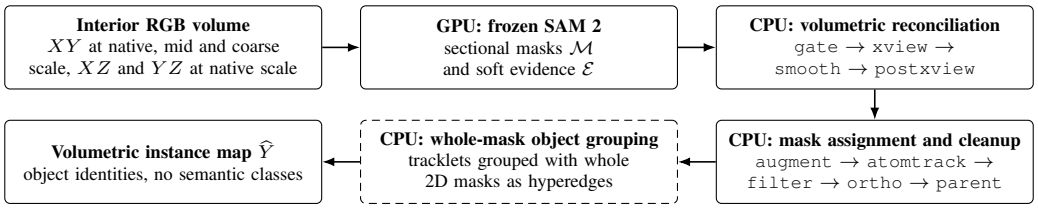



Figure 3: Stage order. The dashed stage is the grouping of Section 3.4. Stage names are labels rather than specifications.

### 3.6 Compared methods

We ran seven existing methods without object-specific prompts, with their official implementations and weights, on the workstation of Appendix F (Table 6). SAM 3 received the generic text prompt "Everything".

## 4 Evaluation design

We used instance F1@0.5 (Stringer et al., 2021) and clDice (Shit et al., 2021) for measurement.

### 4.1 Instance F1@0.5

This measure counts whole objects. Each predicted label is either matched to one reference object or charged as a false positive. A method therefore cannot improve its score by emitting more candidates, and a method that finds an object but splits it in two gains nothing. Its cost is that it needs an exhaustively annotated reference. For a nonempty predicted label $P$ and a nonempty reference object $G$,

$$\mathrm{IoU}(P, G) = |P \cap G|/|P \cup G|. \tag{4}$$

We form the IoU matrix over all nonempty labels on both sides, assign predictions to references one to one by maximizing the total IoU with the Hungarian algorithm, and accept an assigned pair when its IoU is at least 0.5. With $n_P$ predicted labels, $n_G$ reference objects and $TP$ accepted pairs,

$$F_1@0.5 = 2TP/(n_P + n_G). \tag{5}$$

For a three-plane scope, $TP$, $n_P$ and $n_G$ are summed over the three planes before $F_1$ is formed. In the ROI cubes the predicted labels are cropped to the $64^3$ box before they are counted. Average precision at the same threshold, $AP = TP/(n_P + n_G - TP)$, carries the same information because $F_1 = 2AP/(1+AP)$ identically, and we therefore report $F_1$ only. Agreement on merged foreground is not a substitute, because it can be made perfect by a solution that assigns no instances at all.

### 4.2 clDice

This measure compares centerlines. A thin object is penalized by volumetric overlap for an error of one voxel along its whole length, and clDice is insensitive to that while remaining sensitive to a break in the object. Its cost is that it says nothing about how many objects a method produced. With $S(\cdot)$ the skeleton of a mask, against a reference object $G$ and a prediction $P$,

$$\mathrm{tprec} = |S(P) \cap G|/|S(P)|, \quad \mathrm{tsens} = |S(G) \cap P|/|S(G)|, \tag{6}$$

$$\mathrm{clDice} = 2 \cdot \mathrm{tprec} \cdot \frac{\mathrm{tsens}}{\mathrm{tprec} + \mathrm{tsens}}. \tag{7}$$

We report clDice for the single predicted label with the highest IoU against each reference object, and on the public volumes also its centerline sensitivity, tsens. The pixel counts are summed over sections before tprec and tsens are formed. We use 2D skeletons because the voxels are anisotropic. A reference object that no predicted label overlaps scores zero and stays in the mean.

### 4.3 SCORING FRAMEWORK

The three exhaustively annotated cubes give six scored scopes, the three planes pooled and the ROI cube of each (Figure 2). These carry every claim that rests on instance F1@0.5. The qualitative cube and the eleven public volumes have only some objects labeled (positive-only). Precision, instance F1 and AP are therefore undefined on both. On the qualitative cube we report detection counts by semantic category. On the eleven public volumes we report the best IoU that any single predicted label achieves against each reference object. That measure charges nothing for labels placed outside the reference object, and Table 9 gives the number of labels per specimen.

Across the eleven public specimens we test the sign of the per-specimen difference with a two-sided exact binomial test, with the specimen as the unit and ties excluded, and we report each method's median best IoU, the win counts and the $p$ value.

To test whether the ordering is produced by objects small enough that both their annotation and their overlap are uncertain, we removed predicted and reference objects smaller than 20 pixels in plane, judged in the ROI cubes by the largest $XY$ cross-section. This secondary analysis, proposed after the first results were in, is given in full in Table 11.

## 5 RESULTS

### 5.1 INSTANCE-LEVEL PERFORMANCE ON THE RGB VOLUME

On instance F1@0.5, WINZE scores highest in four of the six scopes, three of them on the two independent cubes. On the `DS00007_cube0001` ROI it ranks third behind u-Segment3D and µSAM, and on the `DS00068_cube0002` ROI it ranks second behind Cellpose-SAM, which places one label in that ROI. In both ROIs WINZE matches more reference objects than any compared method (23 and 4), at a lower precision. On mean clDice, WINZE scores highest in all six scopes (Table 2). Over the eleven public specimens, WINZE's median best IoU is 0.668 (Table 3), with a range of 0.120–0.834. Against DEVA it wins nine specimens to two on best IoU and on clDice, and seven to four on centerline sensitivity, without reaching significance. Both losses on best IoU are narrow, 0.834 against 0.856 and 0.731 against 0.771. Figure 6 shows the results on one specimen in 3D.

On the specimen where WINZE scores lowest, a fossil squid beak (`DS00173_7Ob4vJ8`), µSAM and WINZE give the highest best IoU of any method (0.126 and 0.120). WINZE reaches a clDice of 0.230 against 0.006 for DEVA on the same specimen.

The secondary 20-pixel analysis raises WINZE's three-plane scores to 0.314–0.495 without changing the ordering in any three-plane scope (Table 11). In the ROI scopes WINZE moves from third to second on `DS00007_cube0001` and from second to third on `DS00068_cube0002`.

One $256^3$ cube under the frozen configuration takes 33.58 minutes on our workstation, of which 30.86 minutes (91.9%) is detection by SAM 2 (Table 19).

### 5.2 QUALITATIVE GRANULARITY AND A FAILURE REGIME

WINZE produces a finer partition than DEVA in the sectional overlays (Figures 4 and 5), including many small regions in the surrounding rock. Fossils of ammonites and thick bivalve shells, all of which are large and bright, are recovered most often (Table 4). WINZE recovers 6 of 6 ammonites and 2 of 2 thick bivalve shells, with median best IoU 0.869 and 0.878. DEVA is the only compared method to reach two objects in any category, at 2 of 6 ammonites with a median best IoU of 0.290 and 2 of 25 plants at 0.053. Every other compared method recovers at most one object in any category. Cellpose-SAM recovers 1 of 2 thick bivalve shells, with a median best IoU of 0.360.

In contrast, thin bivalve shells are the most difficult category for WINZE, and no compared method recovers any of them. WINZE recovers 1 of 5 thin bivalve shells at median best IoU 0.339, against 0.66–0.88 across its other categories. WINZE recovers 20 of 25 plants. Of the 8 particles, WINZE recovers 6 and DEVA none, at median best IoU 0.674 and 0.000.

Table 2: The quantitative results for seven compared methods and WINZE. (a) Instance F1@0.5 on the three exhaustively annotated cubes. Bold marks the single best method in a column. $n$ is the number of reference objects. † development cube. (b) The mean clDice. Medians and WINZE's count of objects scoring exactly zero are in Table 13.

| Method | DS00007_cube0001† | | DS00068_cube0002 | | DS00136_cube0004 | |
|---|---|---|---|---|---|---|
| | 3 planes ($n$=454) | ROI cube ($n$=57) | 3 planes ($n$=252) | ROI cube ($n$=26) | 3 planes ($n$=339) | ROI cube ($n$=47) |
| *(a) instance F1@0.5* | | | | | | |
| WINZE (ours) | **0.228** | 0.153 | **0.258** | 0.072 | **0.128** | **0.207** |
| DEVA | 0.111 | 0.000 | 0.082 | 0.071 | 0.067 | 0.040 |
| Cellpose-SAM | 0.088 | 0.000 | 0.030 | **0.074** | 0.000 | 0.000 |
| µSAM | 0.113 | 0.156 | 0.027 | 0.057 | 0.004 | 0.035 |
| u-Segment3D | 0.174 | **0.161** | 0.015 | 0.000 | 0.011 | 0.000 |
| SAM 3 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SAMamba3D | 0.000 | 0.000 | 0.000 | 0.000 | 0.006 | 0.040 |
| VISTA3D | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| *(b) mean clDice* | | | | | | |
| WINZE (ours) | **0.485** | **0.491** | **0.418** | **0.359** | **0.326** | **0.444** |
| DEVA | 0.231 | 0.102 | 0.073 | 0.040 | 0.069 | 0.023 |
| Cellpose-SAM | 0.114 | 0.066 | 0.024 | 0.027 | 0.010 | 0.041 |
| µSAM | 0.195 | 0.321 | 0.056 | 0.043 | 0.033 | 0.031 |
| u-Segment3D | 0.309 | 0.324 | 0.017 | 0.042 | 0.008 | 0.003 |
| SAM 3 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SAMamba3D | 0.018 | 0.049 | 0.023 | 0.066 | 0.021 | 0.041 |
| VISTA3D | 0.008 | 0.007 | 0.017 | 0.020 | 0.005 | 0.000 |

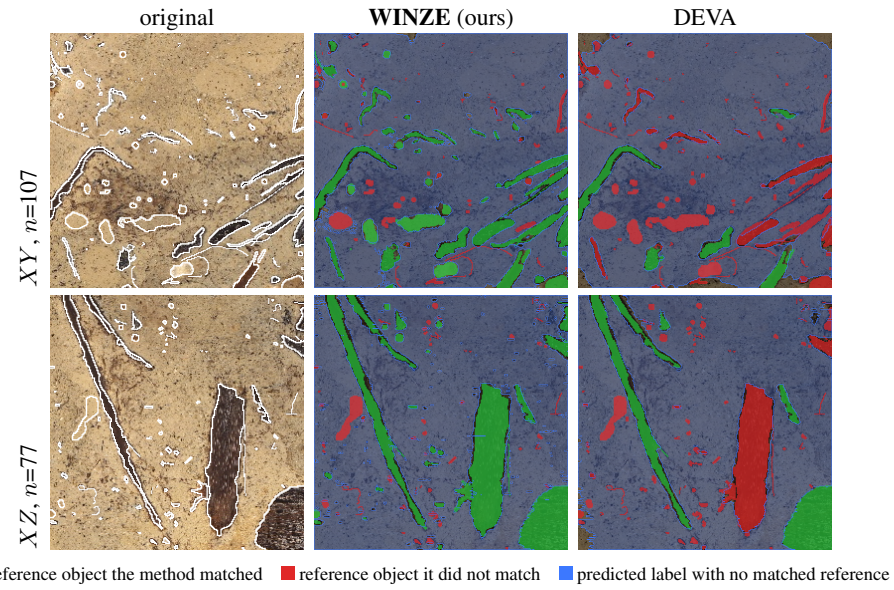



Figure 4: Comparison between the best and the second-best methods on DS00068_cube0002, on the in-plane $XY$ plane and the orthogonal $XZ$ plane. On the three planes of this cube, WINZE ranks first and DEVA second (Table 2). The left column overlays the reference outlines in white. In the WINZE and DEVA columns, a large blue region is one unmatched predicted label that covers the rock matrix. Here $n$ counts reference objects. Figure 5 adds the $YZ$ plane and all compared methods.

Table 3: Results for all eleven public natural-color volumes, with best IoU taken per reference object (Section 4). Wins count the specimens on which WINZE is higher and lower, and $p$ is from the binomial test of Section 4. The clDice and centerline sensitivity columns repeat the test for the best-IoU label. Bold marks the highest median. µSAM failed in an upstream step on one specimen (DS00011 U2ovTAd), which is excluded from its row rather than scored as zero.

| | best IoU | | | clDice | | centerline sens. | |
|---|---|---|---|---|---|---|---|
| Method | median | wins | $p$ | wins | $p$ | wins | $p$ |
| WINZE (ours) | **0.668** | – | – | – | – | – | – |
| DEVA | 0.552 | 9–2 | 0.065 | 9–2 | 0.065 | 7–4 | 0.549 |
| Cellpose-SAM | 0.000 | 11–0 | 0.001 | 11–0 | 0.001 | 11–0 | 0.001 |
| µSAM | 0.143 | 9–1 | 0.021 | 10–0 | 0.002 | 9–1 | 0.021 |
| u-Segment3D | 0.008 | 11–0 | 0.001 | 11–0 | 0.001 | 9–2 | 0.065 |
| SAM 3 | 0.000 | 11–0 | 0.001 | 11–0 | 0.001 | 11–0 | 0.001 |
| SAMamba3D | 0.141 | 11–0 | 0.001 | 11–0 | 0.001 | 6–5 | 1.000 |
| VISTA3D | 0.006 | 11–0 | 0.001 | 11–0 | 0.001 | 11–0 | 0.001 |

Table 4: Objects recovered at IoU $\geq 0.5$ in the qualitative cube, counted by semantic category. Column headings give the number of annotated objects in each category. The categories were defined by the annotator. The outlines were drawn from scratch. Appendix B.1 gives the provenance.

| Method | ammo-nite (6) | beak (4) | plants (25) | thick bivalve shell (2) | leaf (3) | particle (8) | thin bivalve shell (5) | total (53) |
|---|---|---|---|---|---|---|---|---|
| WINZE (ours) | 6 | 3 | 20 | 2 | 2 | 6 | 1 | 40 |
| DEVA | 2 | 1 | 2 | 0 | 1 | 0 | 0 | 6 |
| Cellpose-SAM | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 4 |
| µSAM | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 3 |
| u-Segment3D | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 2 |
| SAM 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| SAMamba3D | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| VISTA3D | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

### 5.3 What the comparison does not isolate

Across the four cubes the grouping stage cuts the fragment count by 35 to 59% (Table 14). We have not audited which individual merges are incorrect. No stage-removal ablation with scores, single-view control or u-Segment3D assembly of the same SAM 2 masks was run. When several predicted labels may jointly match one reference object as a bundle, WINZE has 397 true positives and 81% of its bundles are a single label, with a median bundle size of 1 and a maximum of 21. DEVA has 87% single-label bundles and a maximum of 2, and µSAM 61% and 17 (Table 12).

On grayscale input, WINZE's gap between in-plane and orthogonal F1 moves from 0.344 to 0.450 on the development cube, and from 0.273 to 0.266 and 0.138 to 0.117 on the two independent cubes. On the development cube the gaps of Cellpose-SAM, µSAM and DEVA, which take RGB directly, move by $-0.020$, $+0.020$ and $-0.036$. Five of the six gaps are essentially unchanged, and we claim no color-specific gain. On the development cube, Cellpose-SAM's three-plane score rises from 0.088 to 0.253, above that of WINZE under either input (Table 18).

## 6 Discussion and limitations

**The strong method for reading natural-color tomography.** WINZE scores highest among the eight methods in most of the quantitative scopes and in the qualitative evaluation (Tables 2 and 4). In F1 it achieves scores 1.3 to 3.1 times those of the best compared method on the three sets of scored planes. It leads on one of the ROI cubes by a factor of five, the largest ratio of any scope. On the

eleven public specimens WINZE has a higher best IoU on every specimen against five of the seven methods (Table 3).

The weakness of WINZE on thin bivalve shells (Table 4) likely reflects contrast, because these shells are all bright and the matrix of `DS00173_cube0006` is itself bright. A bright object therefore separates from it by a smaller margin than a dark object (Table 15).

WINZE fixes the accepted foreground and reassigns identity only within it. The same separation underlies u-Segment3D (Zhou et al., 2025). WINZE differs in two respects. It detects with a frozen SAM 2 on RGB sections, and it assigns identity to 2D atoms whose pixels are never altered. We expect the former choice to raise the number of detected objects and the latter to keep sectional boundaries intact through assembly. WINZE has a higher recall than u-Segment3D on every cube (Table 8), and its outputs show fewer assembly artifacts in the sectional overlays (Figure 5). WINZE's F1 lead over u-Segment3D in five of the six scopes (Table 2) is consistent with these two effects.

**Near-zero baselines fail for different reasons.** VISTA3D, SAMamba3D, Cellpose-SAM and SAM 3 score an F1 of exactly zero in half or more of the scopes (Table 2). VISTA3D and Cellpose-SAM label orders of magnitude less of each plane as foreground than WINZE or DEVA (Table 10). They rarely reach the matching stage at all. SAMamba3D fails in the opposite direction. It assigns every voxel to one of four material classes, and its output therefore carries almost no instance boundaries. SAM 3 returns no object at all for the prompt "Everything". Reporting the four together as low-scoring would merge different failures.

The same frozen SAM 2 with its masks linked along $Z$ by overlap alone scores lower than WINZE in all four scopes of the two cubes on which its sectional masks are precise in plane (Table 16). Precise sectional masks alone therefore do not yield a high score.

**Weakness of WINZE and challenges of this domain.** Although WINZE leads on this domain, its absolute scores remain low, and the domain is far from solved. Under the same scorer, Cellpose-SAM and VISTA3D reach much higher values on the domains they were built for (Table 21). Within the $XY$ plane WINZE returns fewer labels than there are reference objects, and on two of the three cubes most of these labels are matched (Table 8). On the orthogonal planes its recall is largely kept. The number of labels grows to several times the number of reference objects, and precision collapses. Table 17 shows that part of the surplus consists of fragments of a single reference object. The rest lies largely in the surrounding rock, which is not held as one label across sections either, and the atoms there have no stable label to vote for. These results indicate that WINZE keeps its recall across planes but does not always hold each object as one label across sections.

An approach that assumes spatial continuity may resolve this. WINZE already links atoms between adjacent sections by rule. Such an approach would instead impose continuity on the volume as a whole. Building a property of the domain into the method has worked elsewhere. FlowSAM adds optical flow to SAM and was reported to outperform earlier methods in video object segmentation (Xie et al., 2024). SAI3D uses geometric priors to merge SAM masks into instances in 3D scenes (Yin et al., 2024). Natural-color volumes may likewise benefit from methods that take their own properties as prior assumptions.

## 7 Conclusion

WINZE, an instance segmentation method designed for natural-color volumes, scores highest in most scopes against seven segment-everything methods on 1,239 manually annotated instances. However, its absolute scores remain low. WINZE sometimes fails to hold each object as one label across sections. Methods that build the spatial continuity of objects into their assumptions are a natural next step for natural-color tomography.

### Ethics statement

The application concerns images of rock material rather than clinical diagnosis. Erroneous instance boundaries can nevertheless propagate into scientific measurements and anatomical interpretations, and predictions should not be treated as established structures without validation. All specimens

obtained for this study were collected with permission from the relevant national and regional authorities. The institutions are withheld to preserve anonymity and will be named in the camera-ready version. The annotations and their recorded times are the work of one author, and no other human participant took part.

### REPRODUCIBILITY STATEMENT

The supplementary material contains, for one scored cube (`DS00068_cube0002`), the RGB images, metadata, reference labels and label volumes predicted by every method, with the license of the cube data. The source code of WINZE will be released upon acceptance. Section 3.4 describes the grouping stage, which runs separately from the driver. The eleven public volumes are available from Ikegami et al. (2025). Section 3 gives the stage order and the backbone checkpoint, and Appendices C.1 and C.2 give every rule and threshold. No weights were trained or fine-tuned, and no threshold was set from the annotations. Table 6 identifies each compared method by version or commit, weights and invocation. Section 4 defines the matching rule, Appendix B.1 describes how the cubes, planes and ROIs were selected, and Tables 7, 11 and 12 give the per-plane scores, the secondary 20-pixel analysis and the bundle analysis. Appendix F and Table 19 give the computing environment and the measured cost.

### AI USE STATEMENT

We used LLM assistants in this work. The assistants used were Claude (Anthropic) and GPT (OpenAI).

***Research conception.*** The problem addressed here, that natural-color tomographic volumes are not served by existing segmentation methods and that a method for them was worth developing, was identified by the authors. The research plan and the design choices of the method, namely (a) reading sections in three orthogonal orientations and the $XY$ sections at several scales and (b) retaining pre-threshold evidence rather than binarized masks and deferring identity assignment to a rule-based volumetric stage, were likewise made by the authors. No assistant contributed to the problem formulation or to these choices.

***Implementation.*** The implementation of the method, the evaluation harness, and the analysis scripts were written by these assistants under author direction. This includes the reimplementation, from the published description, of the aggregation of u-Segment3D and an adaptation of its label diffusion, which replaced the u-Segment3D code as a dependency. The per-object loop of the smoothing stage, its box margin of 25 voxels and its value of 0.7 for the original labels follow the u-Segment3D implementation.

***Research design and analysis.*** Within the plan above, assistants participated in the design of individual experiments and in hypothesis formulation, including the grayscale ablation reported in Table 18 and the selection rule used to place the scored planes and the three-dimensional regions of interest (Appendix B.1). For the stages that implement choices (a) and (b), assistants also proposed rules and thresholds, including the assignment of unaltered 2D masks to tracks and the track acceptance conditions. The authors set the goals, evaluated each proposal on the data, and decided what to adopt before the configuration was frozen. The assistants executed the scoring runs, aggregated the metrics, and produced the figures and tables from the saved predictions. They also proposed interpretations of the results, which the authors checked against the saved predictions and then accepted, revised or rejected.

***Data.*** The annotations were produced manually. Scripts written by assistants were used only to rename and move files and to make the edits to the label volumes listed next, and the authors confirmed that the image content was unchanged. Instance indices were compacted, fifteen voxels of overlap in one cube were resolved by a majority vote over exclusive neighbors, the segments of one cube were renamed, and one segment was deleted.

***Literature survey and writing.*** The assistants retrieved and summarized related work, drafted and revised the manuscript text, and translated text written by the authors into English.

***Tasks without assistance.*** We have not used generative AI tools to generate synthetic data sets or to develop theoretical models. Formulating mathematical claims, proving them, writing proofs, and qualitative or thematic data analysis are not applicable to this work.

***Verification.*** We have reviewed all AI-assisted work. The authors checked the final manuscript text. Every reference was checked against a primary source by the authors. All annotation labels were inspected visually by the annotating author, and automated checks for out-of-plane voxels, overlaps, empty segments, and index gaps were run after every save. Label volumes were verified by round-trip export and re-import.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Anwai Archit, Luca Freckmann, Sushmita Nair, Nabeel Khalid, Paul Hilt, Vikas Rajashekar, Marei Freitag, Carolin Teuber, Melanie Spitzner, Constanza Tapia Contreras, Genevieve Buckley, Sebastian von Haaren, Sagnik Gupta, Marian Grade, Matthias Wirth, Günter Schneider, Andreas Dengel, Sheraz Ahmed, and Constantin Pape. Segment Anything for microscopy. *Nature Methods*, 22(3):579–591, 2025. doi: 10.1038/s41592-024-02580-4.

John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. doi: 10.1109/TPAMI.1986.4767851.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Dídac Surís, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, François Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1316–1326, 2023. doi: 10.1109/iccv51070.2023.00127.

Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, John Buatti, Stephen Aylward, James V. Miller, Steve Pieper, and Ron Kikinis. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging*, 30(9):1323–1341, 2012. doi: 10.1016/j.mri.2012.05.001.

Karol Gotkowski, Shuvam Gupta, Jose R.A. Godinho, Camila G.S. Tochtrop, Klaus H. Maier-Hein, and Fabian Isensee. ParticleSeg3D: A scalable out-of-the-box deep learning segmentation solution for individual particle characterization from micro CT images in mineral processing and recycling. *Powder Technology*, 434:119286, 2024. doi: 10.1016/j.powtec.2023.119286.

Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Daguang Xu, and Wenqi Li. A short review and evaluation of SAM2's performance in 3D CT image segmentation. *arXiv preprint arXiv:2408.11210*, 2024. doi: 10.48550/arXiv.2408.11210.

Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, Daguang Xu, and Wenqi Li. VISTA3D: A unified segmentation foundation model for 3D medical imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20863–20873, 2025. doi: 10.1109/cvpr52734.2025.01943.

Yasuhiro Iba, Aya Kubota, Yusuke Takeda, Mehmet Oguz Derin, Shin Ikegami, Jörg Mutterlose, Takahiro Harada, Tomonori Takeuchi, and Kazuki Tainaka. Nature visible only digitally. *Patterns*, 6(4):101210, 2025. doi: 10.1016/j.patter.2025.101210.

Shin Ikegami, Yusuke Takeda, Jörg Mutterlose, and Yasuhiro Iba. Fossil lower beaks of Cretaceous squids. Figshare+, 2025. doi: 10.25452/figshare.plus.25736631.

Shin Ikegami, Jörg Mutterlose, Kanta Sugiura, Yusuke Takeda, Mehmet Oguz Derin, Aya Kubota, Kazuki Tainaka, Takahiro Harada, Harufumi Nishida, and Yasuhiro Iba. Earliest octopuses were giant top predators in Cretaceous oceans. *Science*, 392(6796):406–410, 2026a. doi: 10.1126/science.aea6285.

Shin Ikegami, Kanta Sugiura, Mehmet Oguz Derin, Yusuke Takeda, Jörg Mutterlose, Shintaro Sasaki, Neil H. Landman, Kazuki Tainaka, Aya Kubota, Harufumi Nishida, and Yasuhiro Iba. Squids radiated globally prior to the Cretaceous–Paleogene mass extinction. *Geology*, 54(9): 938–942, 2026b. doi: 10.1130/G54372.1.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, 2023. doi: 10.1109/iccv51070.2023.00371.

Hans Liebl, David Schinz, Anjany Sekuboyina, Luca Malagutti, Maximilian T. Löffler, Amirhossein Bayat, Malek El Husseini, Giles Tetteh, Katharina Grau, Eva Niederreiter, Thomas Baum, Benedikt Wiestler, Bjoern Menze, Rickmer Braren, Claus Zimmer, and Jan S. Kirschke. A computed tomography vertebral segmentation dataset with anatomical variations and multi-vendor scanner data. *Scientific Data*, 8(1):284, 2021. doi: 10.1038/s41597-021-01060-0.

Zudi Lin, Donglai Wei, Mariela D. Petkova, Yuelong Wu, Zergham Ahmed, Krishna Swaroop K, Silin Zou, Nils Wendt, Jonathan Boulanger-Weill, Xueying Wang, Nagaraju Dhanyasi, Ignacio Arganda-Carreras, Florian Engert, Jeff Lichtman, and Hanspeter Pfister. NucMM dataset: 3D neuronal nuclei instance segmentation at sub-cubic millimeter scale. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 164–174, 2021. doi: 10.1007/978-3-030-87193-2_16.

Maximilian T. Löffler, Anjany Sekuboyina, Alina Jacob, Anna-Lena Grau, Andreas Scharr, Malek El Husseini, Mareike Kallweit, Claus Zimmer, Thomas Baum, and Jan S. Kirschke. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2(4):e190138, 2020. doi: 10.1148/ryai.2020190138.

Miquel López Escoriza and Pau Amargant Álvarez. Automatic segmentation of 3D CT scans with SAM2 using a zero-shot approach. *arXiv preprint arXiv:2603.23116*, 2026. doi: 10.48550/arXiv.2603.23116.

Akshay Mehra, Bolton Howes, Ryan Manzuk, Alex Spatzier, Bradley M. Samuels, and Adam C. Maloof. A novel technique for producing three-dimensional data using serial sectioning and semi-automatic image classification. *Microscopy and Microanalysis*, 28(6):2020–2035, 2022. doi: 10.1017/S1431927622012442.

Satoru Muro, Takuya Ibara, Yuzuki Sugiyama, Akimoto Nimura, and Keiichi Akita. Evaluating accuracy in artificial intelligence-powered serial segmentation for sectional images applied to morphological studies with three-dimensional reconstruction. *Microscopy*, 74(2):107–116, 2025. doi: 10.1093/jmicro/dfae054.

Phuc Nguyen, Minh Luu, Anh Tran, Cuong Pham, and Khoi Nguyen. Any3DIS: Class-agnostic 3D instance segmentation by 2D mask tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3636–3645, 2025. doi: 10.1109/cvpr52734.2025.00344.

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)*, pp. 4018–4028, 2024. doi: 10.1109/cvpr52733.2024.00385.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

Marius Pachitariu, Michael Rariden, and Carsen Stringer. Cellpose-SAM: superhuman generalization for cellular segmentation. *bioRxiv*, 2025. doi: 10.1101/2025.04.28.651001.

Indu Panigrahi, Ryan A. Manzuk, Adam C. Maloof, and Ruth Fong. Improving data-efficient fossil segmentation via model editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4829–4838, 2023. doi: 10.1109/CVPRW59228.2023.00511.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. doi: 10.48550/arXiv.1704.00675.

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. In *International Conference on Learning Representations (ICLR)*, 2025.

Anjany Sekuboyina, Malek E. Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, Martin Urschler, Maodong Chen, Dalong Cheng, Nikolas Lessmann, Yujin Hu, Tianfu Wang, Dong Yang, Daguang Xu, Felix Ambellan, Tamaz Amiranashvili, Moritz Ehlke, Hans Lamecker, Sebastian Lehnert, Marilia Lirio, Nicolás Pérez de Olaguer, Heiko Ramm, Manish Sahu, Alexander Tack, Stefan Zachow, Tao Jiang, Xinjun Ma, Christoph Angerman, Xin Wang, Kevin Brown, Alexandre Kirszenberg, Élodie Puybareau, Di Chen, Yiwei Bai, Brandon H. Rapazzo, Timyoas Yeah, Amber Zhang, Shangliang Xu, Feng Hou, Zhiqiang He, Chan Zeng, Zheng Xiangshang, Xu Liming, Tucker J. Netherton, Raymond P. Mumme, Laurence E. Court, Zixun Huang, Chenhang He, Li-Wen Wang, Sai Ho Ling, Lê Duy Huỳnh, Nicolas Boutry, Roman Jakubicek, Jiri Chmelik, Supriti Mulay, Mohanasankar Sivaprakasam, Johannes C. Paetzold, Suprosanna Shit, Ivan Ezhov, Benedikt Wiestler, Ben Glocker, Alexander Valentinitsch, Markus Rempfler, Björn H. Menze, and Jan S. Kirschke. VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images. *Medical Image Analysis*, 73:102166, 2021. doi: 10.1016/j.media.2021.102166.

Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. clDice – a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16555–16564, 2021. doi: 10.1109/cvpr46437.2021.01629.

Victor Spitzer, Michael J. Ackerman, Ann L. Scherzinger, and David Whitlock. The Visible Human male: A technical report. *Journal of the American Medical Informatics Association*, 3(2):118–130, 1996. doi: 10.1136/jamia.1996.96236280.

Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, 2021. doi: 10.1038/s41592-020-01018-x.

Qing Su, Kaiyang Li, Yuan Zhuang, Fei Miao, and Shihao Ji. Open-world video segmentation. *arXiv preprint arXiv:2606.15632*, 2026. doi: 10.48550/arXiv.2606.15632.

Kanta Sugiura, Shin Ikegami, Yusuke Takeda, Jörg Mutterlose, Mehmet Oguz Derin, Aya Kubota, Harufumi Nishida, Kazuki Tainaka, Takahiro Harada, Neil H. Landman, and Yasuhiro Iba. The oldest sepioid cephalopod from the Cretaceous discovered by Digital fossil-mining with zero-shot learning AI. *Communications Biology*, 9(1):301, 2026. doi: 10.1038/s42003-026-09519-9.

Donglai Wei, Zudi Lin, Daniel Franco-Barranco, Nils Wendt, Xingyu Liu, Wenjie Yin, Xin Huang, Aarush Gupta, Won-Dong Jang, Xueying Wang, Ignacio Arganda-Carreras, Jeff W. Lichtman, and Hanspeter Pfister. MitoEM dataset: Large-scale 3D mitochondria instance segmentation from EM images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 66–76, 2020. doi: 10.1007/978-3-030-59722-1_7.

Adrian Wolny, Lorenzo Cerrone, Athul Vijayan, Rachele Tofanelli, Amaya Vilches Barro, Marion Louveaux, Christian Wenzl, Sören Strauss, David Wilson-Sánchez, Rena Lymbouridou, Susanne S. Steigleder, Constantin Pape, Alberto Bailoni, Salva Duran-Nebreda, George W. Bassel, Jan U. Lohmann, Miltos Tsiantis, Fred A. Hamprecht, Kay Schneitz, Alexis Maizel, and Anna Kreshuk. Accurate and versatile 3D segmentation of plant tissues at cellular resolution. *eLife*, 9: e57613, 2020. doi: 10.7554/elife.57613.

Junyu Xie, Charig Yang, Weidi Xie, and Andrew Zisserman. Moving object segmentation: All you need is SAM (and flow). In *Asian Conference on Computer Vision (ACCV)*, pp. 291–308, 2024. doi: 10.1007/978-981-96-0972-7_17.

Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. MaskClustering: View consensus based mask graph clustering for open-vocabulary 3D instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28274–28284, 2024. doi: 10.1109/cvpr52733.2024.02671.

Yeqing Yang, Le Xu, and Lixia Tian. SAM2-3dMed: Empowering SAM2 for 3D medical image segmentation. *arXiv preprint arXiv:2510.08967*, 2025. doi: 10.48550/arXiv.2510.08967.

Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. SAI3D: Segment any instance in 3D scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3292–3302, 2024. doi: 10.1109/cvpr52733.2024.00317.

Rui Zhang, Xianzhi Song, Linqi Zhu, Branko Bijeljic, Gensheng Li, and Martin J. Blunt. SAMamba3D: Adapting Segment Anything for generalizable three-dimensional segmentation of multiphase pore-scale images. *Advances in Geo-Energy Research*, 21(2):109–124, 2026. doi: 10.46690/ager.2026.08.03.

Yibo Zhao, Yigong Zhang, and Jin Xie. MV3DIS: Multi-view mask matching via 3D guides for zero-shot 3D instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Felix Y. Zhou, Zach Marin, Clarence Yapp, Qiongjing Zou, Benjamin A. Nanes, Stephan Daetwyler, Andrew R. Jamieson, Md Torikul Islam, Edward Jenkins, Gabriel M. Gihana, Jinlong Lin, Hazel M. Borges, Bo-Jui Chang, Andrew Weems, Sean J. Morrison, Peter K. Sorger, Reto Fiolka, Kevin M. Dean, and Gaudenz Danuser. Universal consensus 3D segmentation of cells from 2D segmented stacks. *Nature Methods*, 22(11):2386–2399, 2025. doi: 10.1038/s41592-025-02887-w.

# A Related work and the comparison boundary

**Foundation models.** Segment Anything Model (SAM) established high-performance image segmentation (Kirillov et al., 2023), and SAM 2 extended it to video (Ravi et al., 2025). These models achieved strong label generation. Segment-everything, however, has been defined only for single images. SAM 3 moves the prompt up to the concept level (Carion et al., 2026). The structures in our volumes, however, are outside the vocabulary these models were built around because natural-color tomography has almost no public data. We therefore expected SAM 3 to transfer poorly and ran it with the generic text prompt "Everything" (Table 6). We also tried other prompts in video mode on the first 32 sections of `DS00007_cube0001`, `DS00173_cube0006` or both. The prompts "Object", "rock", "stone", "fossil" and "Black object" returned no object, and "White object" returned one. Lowering the output threshold from 0.5 to 0 on `DS00007_cube0001` still left "Everything" without any object.

**Video object segmentation.** The decoupled video segmentation approach (DEVA) separates task-specific image segmentation from class-agnostic bidirectional propagation and consolidates detections across frames (Cheng et al., 2023). Reading the temporal axis as a spatial one turns such a tracker into a volumetric segmenter (López Escoriza & Amargant Álvarez, 2026). Sugiura et al. (2026) and later work took this route and applied the DEVA algorithm to natural-color tomography, which gave a pseudo segment-everything for volumes (Ikegami et al., 2026a;b). Applying a video tracker to a volume raises problems that belong to the approach rather than to any one implementation. Dependencies between video frames run in one direction of time, whereas dependencies between sections follow the spatial continuity of the objects and hold in both directions (Yang et al., 2025). In CT, video object segmentation produces false positives where an object disappears, and negative clicks on one section do not suppress the false positives that begin on the next (He et al., 2024). Object identity also fragments for a related reason, because object admission rests on frame-wise association rather than explicit validation (Su et al., 2026). Identity in a volume is therefore better decided from agreement between orientations than from scan order. In our comparison DEVA and SAM 3 follow this route, and Table 16 adds a simpler control that links the sectional masks of the SAM 2 checkpoint of WINZE by overlap alone.

**Multi-view scene segmentation.** Open3DIS associates image masks with superpoints and combines image-guided and native 3D proposals (Nguyen et al., 2024). MaskClustering constructs a mask graph using view consensus (Yan et al., 2024). More recent work uses SAM 2 mask tracking with 3D guidance (Any3DIS) or coarse 3D segments to match masks across views (MV3DIS) (Nguyen et al., 2025; Zhao et al., 2026). These methods establish that evidence aggregation and persistent object identity are active, well-developed ideas. Their RGB-D/point-cloud formulations involve camera projection and surface visibility, which differ from interior reslicing. We therefore treat these methods as background rather than as baselines on our volumes.

**Task-specific volume segmentation.** To segment everything within specific domains, many methods train a model for their targets or fine-tune a foundation model. µSAM fine-tunes SAM for microscopy and supports automatic and interactive use (Archit et al., 2025). Cellpose-SAM adapts a SAM encoder for cellular segmentation (Pachitariu et al., 2025). VISTA3D targets automatic and interactive medical volumes (He et al., 2025). SAMamba3D keeps a largely frozen SAM encoder and adds volumetric context for pore-scale volumes (Zhang et al., 2026). Their training domains are grayscale CT and stained or fluorescent RGB microscopy. None of them is designed for natural-color interior volumes. Their scores on our data therefore measure transfer rather than the methods themselves (Table 1).

**Segmentation for serial-sectional images.** Serial-sectional images in natural color also exist in anatomy, such as the Visible Human Project (Spitzer et al., 1996). Muro et al. (2025) evaluated the SAM-Track model for segmenting anatomical and histological structures on such sections. This study targets named structures. WINZE instead returns every object as an instance without a class vocabulary. The volumes in our study resemble those acquired by serial-grinding systems such as GIRI (Mehra et al., 2022). Two studies have segmented such data automatically or semi-automatically (Mehra et al., 2022; Panigrahi et al., 2023). These methods are trained on labels

annotated for predefined classes and cannot segment objects outside them. We therefore treat them as background rather than segment-everything baselines.

**Consensus within volumes.** The closest prior work is u-Segment3D (Zhou et al., 2025). It turns 2D segmentations into gradient fields, combines orthogonal stacks, and recovers instances with gradient dynamics and connected components. Label diffusion and intensity-guided refinement are available on top. Keeping soft evidence, consulting several orientations, and using image appearance therefore do not separate our method from this prior. Two of its properties matter for our input. Its default 2D segmenter is Cellpose, which was trained on cells and struggles with the dark elongated clasts and thin walls our specimens contain in quantity (Table 15). Our voxels are also anisotropic, with a Z-to-XY ratio between 0.37 and 1.56 across the four cubes. The orthogonal sections that a consensus scheme depends on are the ones where a segmenter trained on isotropic images degrades, and those errors accumulate into slit-shaped and plate-shaped artifacts in the reconstructed labels (Figure 5). WINZE re-implements the aggregation of u-Segment3D but uses its result only to guide the labeling and linking of atoms, and the output of WINZE keeps the pixels of the atoms (Section 3).

# B Material details and annotation

Table 5 lists the eleven public volumes, and this appendix describes how the four cubes were selected and annotated.

Table 5: The eleven public natural-color volumes used for replication across specimens, from Ikegami et al. (2025). Each carries one labeled fossil beak and nothing else. The reference is therefore positive-only, and only per-object best IoU and the skeleton measures are defined on it. The labels were made by the original authors with interactive segmentation software, one object at a time, and are not derived from the output of any automatic method. The volumes differ in size by more than two orders of magnitude, from 0.7 to 192.6 megavoxels.

| Specimen ID | sections | Mvoxels |
|---|---|---|
| `DS00009 6SkvYgx` | 42 | 5.2 |
| `DS00011 U2ovTAd` | 38 | 0.7 |
| `DS00034 h0A3Z03` | 39 | 2.3 |
| `DS00041 hs6nJ9o` | 79 | 18.1 |
| `DS00041 qf2yg91` | 75 | 22.2 |
| `DS00081 nsA03VV` | 28 | 2.9 |
| `DS00084 XCCa3MD` | 60 | 20.2 |
| `DS00090 AMpLdq5` | 40 | 7.4 |
| `DS00140 IWPZJL0` | 197 | 192.6 |
| `DS00173 7Ob4vJ8` | 23 | 8.1 |
| `DS00173 bIJ7zso` | 37 | 30.2 |
| total | 658 | 309.9 |

## B.1 Annotation provenance

**Selection of cubes, planes and ROIs.** `DS00007_cube0001` is the cube on which the configuration was developed. The two other quantitative cubes were drawn with a fixed random seed, one from two candidate cubes of each specimen, so that the choice could not follow any property related to the results. The planes were chosen within candidate ranges fixed in advance, $z$ 32–48 for $XY$ and $y$ and $x$ 112–144 for $XZ$ and $YZ$. For each section we computed the fraction of dark pixels, whose luminance is below the Otsu threshold (Otsu, 1979) of the whole volume, and the edge density, the fraction of pixels whose Sobel gradient exceeds the Otsu threshold computed over all $XY$ sections. We took the section closest to the medians of its range, with each measure normalized by its interquartile range. The ROI was chosen by the same rule among $64^3$ windows on a 32-voxel grid that do not overlap the scored planes. On `DS00068_cube0002` the ROI was chosen before

the $YZ$ plane was added to the design, and the scored $YZ$ plane at $x = 124$ passes through it. Only the raw images were used, and no prediction entered these choices.

**Manual segmentation.** The annotation kits contained no predicted labels. Inter-annotator agreement has not been measured. The fourth cube, `DS00173_cube0006`, is annotated positive-only. The voxels are anisotropic and differ between the cubes, namely $30.6 \times 30.6 \times 20$ µm for `DS00007_cube0001`, $52.57 \times 52.57 \times 50$ µm for `DS00068_cube0002`, $27.14 \times 27.14 \times 10$ µm for `DS00136_cube0004`, and $32.03 \times 32.03 \times 50$ µm for `DS00173_cube0006`. Every object in the qualitative cube carries a semantic name, and each was additionally assigned to one stratum on each of three axes, namely color (bright, dark), shape (compact, elongated, thin) and size (large, medium, small). The three groupings each partition the same 53 objects. The assignment was made by the annotator from the reference outlines alone, without reference to any method's output.

### B.1.1 Annotation effort

Painting a 2D instance takes 16 to 20 seconds. The $64^3$ region of `DS00007_cube0001` took 7 to 11 seconds per object-section and 2.2 to 3.4 minutes per object. The recorded painting times total about 10.5 hours for the quantitative references and about 25 hours for all 1,228 instances.

## C Method details

### C.1 Stage rules and thresholds

This appendix states every transition rule of the two CPU phases of Equations 2 and 3, together with the grouping rule of Section 3.4. The main text explains only `gate`, `augment`, `atomtrack` and the grouping. The numeric values that the stage names leave open are in Appendix C.2.

### C.1.1 Phase one: reconciliation

**`gate`** A hysteresis foreground gate (Canny, 1986) is built from the soft evidence and passed to the aggregation as a precomputed binary mask. A high-confidence seed is grown through weaker but connected evidence rather than cut at a single threshold. Settings are `GATE=on`, `FUSE=top2`, `SELFCAL=on`. The gate decides which voxels enter the provisional labeling $U$. The aggregation re-implements the indirect method of u-Segment3D (Zhou et al., 2025) from its published description. It sums the unit gradients of the distance transforms of the three per-view labelings, smooths the field with a Gaussian of $\sigma = 1$ voxel and renormalizes it, moves every gated voxel 40 steps of one voxel, and groups the end points into 26-connected components after a closing of one voxel. It departs from the published method in not smoothing the per-view gradients before summing them, in moving points for 40 steps without momentum where the published method uses 250 steps with momentum, and in grouping the end points by occupancy and a closing rather than by thresholding a smoothed point density. Within the driver, $U$ decides only which label each atom votes for in `atomtrack`, and the output keeps the pixels of the atoms. The grouping uses $U$ again as a weak prior for tracklets.

**`xview`** One-voxel gaps in the provisional three-dimensional labeling are filled where the per-view detections support them. A background voxel is filled when its two neighbors along the scanned axis carry the same positive label and at least `MIN_VIEWS` of the three per-view detections mark it as foreground. The axes $z$, $y$, $x$ are scanned in that order within a pass, and up to `GAP_MAX` passes are run, stopping early when a pass fills nothing. `GAP_MAX` is a pass count and not a gap width. The frozen driver uses `MIN_VIEWS=1` and `GAP_MAX=3`, and the $XY$ view counts toward `MIN_VIEWS`.

**`smooth`** Label diffusion runs for 15 iterations with step size $\alpha = 0.35$ (`REFINE_ITERS=15`, `REFINE_ALPHA=0.35`). The guide $g$ is a single-channel volume formed as the per-voxel mean of the three color channels, not the RGB volume. The guide is rescaled to its 2 to 99.8 percentile range. Each label diffuses within its bounding box enlarged by 25 voxels. In every iteration the value $u$ of a voxel becomes $(1 - \alpha)u + \alpha\bar{u}$, where $\bar{u}$ is the mean of $u$ over the six neighbors weighted by $\exp(-\Delta g^2/(2\sigma^2))$ and $\sigma$ is the mean neighbor difference of the guide within that box. The original

region is then held at $0.7$ or above (`REFINE_CLAMP`). A voxel takes the label when its diffused value exceeds $0.5$, and a later label overwrites an earlier one. The per-object loop, the enlarged bounding box and the value $0.7$ follow the implementation of u-Segment3D (Zhou et al., 2025), and the heat-kernel form of the weight follows the published description. The update differs. The u-Segment3D implementation mixes this weight with a uniform spatial weight in proportion to its own $\alpha$ and mixes the original labels back in every iteration with weight $0.7$, whereas here the weight is the intensity term alone, $\alpha$ is a step size and $0.7$ is a lower bound on the original region.

**`postxview`** The `xview` stage is repeated after smoothing with the same settings.

### C.1.2 Phase two: assignment

**`augment`** Atoms from the three scales are merged under the priority native $>$ mid $>$ coarse. Each lower level contributes only the remainder left uncovered by the levels above it, added as new atoms carrying their provenance. Large matrix fragments are retained at this stage and are not removed as background. A mask of at least the minimum mask size of its level contributes its uncovered remainder when that remainder reaches both the minimum piece of that level and $30\%$ of the mask. Each 4-connected component of the remainder that reaches the minimum component of that level is added as its own atom, and smaller components are discarded. On the cubes reported here the coarse pass added no atom, because its masks are identical to the native masks, and the mid pass added atoms mainly on `DS00007_cube0001`. The coarse masks still enter the grouping of Section 3.4 as $XY$ hyperedges.

**`atomtrack`** Every two-dimensional atom receives a three-dimensional track identifier in one forward pass over $Z$. Write $c_1$ and $c_2$ for the pixel counts an atom shares with its best and second-best candidate label in the labeling from `postxview`, $s$ for the atom's area, and $f$ for the atom's total overlap with foreground.

1. **Voting.** The atom takes its best candidate label when $c_1/s \geq 0.50$, $c_1/f \geq 0.70$, and $c_1 \geq 1.5 \max(c_2, 1)$. An atom at or above the large-atom size is also accepted when $c_1/s \geq 0.35$ and $c_1/f \geq 0.85$, without the runner-up condition. Atoms are processed from the largest. A designated matrix identifier is barred from acting as a positive linker. The frozen configuration sets that identifier to zero, and in the reported runs the guard excludes only the background label.
2. **Inheritance.** An atom not assigned by voting inherits the final track of the atom at $z - 1$ with the highest intersection over union among those with intersection over union at least $0.20$ and a ratio of the larger to the smaller area of at most $2.85$. Inheritance is refused as ambiguous when the two best candidates differ in intersection over union by less than $0.15$ and belong to different tracks. If the best track is already taken in the section, the atom does not fall back to the second candidate.
3. **New tracks.** Atoms left over start new track identifiers. Atoms rejected by the gate are kept rather than dropped.

One atom per track per section is enforced. An atom that loses a contested label goes on to inheritance and starts a new track only when inheritance also fails.

**`filter`** Per-track features are aggregated, namely foreground confidence and local contrast, and written to a track table. The orthogonal-view support is added by the next stage.

The four acceptance conditions are stated in Appendix C.2. Rejected tracks smaller than `V_MIN_LOG` ($50$ voxels on the cubes) are not written to the table and therefore cannot be restored by `ortho`.

**`ortho`** A voxel counts as orthogonally supported when the cross-view mask covering it continues to $z + 1$ or $z - 1$ at the same $(y, x)$. Support is computed as $S = \mathrm{cont}(XZ) \vee \mathrm{cont}(YZ)$, where $\mathrm{cont}(L)$ holds where $L[z] = L[z \pm 1]$ and $L > 0$ along $Z$. The per-track voxel-weighted mean of $S$ is appended to the track table and rescues tracks that the filter rejected, at the thresholds in Appendix C.2. This is direct cross-section evidence of extent, which in-plane overlap alone does not provide.

**`parent`** Fragment tracks are grouped using the coarse full-frame masks as context. In each section, every atom lying at least $50\%$ inside a coarse mask votes its track into the group of the track that holds the remainder atoms `augment` added from that mask. A child track joins its best parent once the accumulated vote covers at least $50\%$ of the child's voxels. Chains are compressed. Voxels are recolored by parent, and tracks that are not children keep their own identifier. A coarse mask enters this stage only at a fixed minimum area of $20{,}000$ native pixels or more, which is $31\%$ of a $256 \times 256$ section. On the four cubes reported here the coarse pass added no atom, and the stage therefore grouped no track.

### C.1.3 WHOLE-MASK OBJECT GROUPING

**Tracklets** Before grouping, all `augment` atoms are linked along $Z$, and the resulting tracklets are then cut to the accepted foreground. A tracklet can therefore skip a section in which its linking atom lies outside the foreground. Two atoms in adjacent sections are eligible when their intersection over union is at least $0.20$ and their area ratio at most $2.85$. The score of a link is the intersection over union, plus $0.10$ when the $U$ label covering at least half of each atom is the same for both atoms, minus $0.10$ times the mean absolute difference of their mean RGB values divided by $255$. A link is kept when the two atoms are each other's best and each leads its runner-up by at least $0.15$. Each tracklet is one fragment of the grouping.

A pair of fragments is accepted when all positive conditions below hold and no negative condition holds. A view is informative for a pair when its observation rate and projection rate reach their thresholds.

| | |
|---|---|
| agreement across views | $\geq 0.70$ |
| observation rate | $\geq 0.50$ |
| projection rate | $\geq 0.20$ |
| RGB histogram similarity | $\geq 0.70$ over 16 bins per channel |
| local membership (pair proposal) | $\geq 0.50$ |
| supporting views | $\geq 2$ |
| negative evidence | agreement $\leq 0.20$ in at least two informative views<br>or color similarity $< 0.25$ |

Within a mask, a fragment is a contributor when at least half of its pixels in that section, and at least four pixels, lie in the mask. The largest contributor is the anchor, and candidate pairs join the anchor with every other contributor holding at least $1\%$ of the anchor's pixels. Agreement and observation are ratios of inner products, and projection is the inner product itself, between square-root-normalized distributions of each fragment over masks, observed sections and all sections. Color similarity is the inner product of the square-root-normalized 16-bin histograms of the three channels. Accepted pairs are merged greedily in decreasing order of strength, the second-highest agreement among informative views times the color similarity. A merge is also refused when the largest fragments of the two groups have color similarity below $0.70$, and a vetoed pair blocks every later merge that would join its two groups.

### C.1.4 DETECTION SETTINGS

All four detection passes run on RGB input in bfloat16, over all sections. Points per batch is $256$ and a single full crop is used (`crop_n_layers`$= 0$). Mask rejection uses the unmodified filters of the base generator at pred_iou_thresh $= 0.55$ and stability_score_thresh $= 0.62$. The model is built with `apply_postprocessing=False`, TF32 is enabled on the GPU, and the other generator settings, including box non-maximum suppression at $0.7$, are the defaults.

| pass | input | field of view | points per side | area exclusion |
|---|---|---|---|---|
| native $XY$ | original | in-plane tiles | 64 | 0.20 |
| mid $XY$ | in-plane $0.5\times$ | in-plane tiles | 64 | 0.85 |
| cross $XZ, YZ$ | original | whole section | 32 | 0.20 |
| coarse $XY$ | original | whole field, $\leq 1024$ px | 64 | 0.85 |

Table 20 gives a post-hoc sweep of the points per side of the native pass. Only the evidence of the native $XY$ pass and of the cross pass enters `gate`. The mid and coarse passes contribute masks

only, and their area exclusion has no effect on the output. The area exclusion column is the fraction of the whole section, or of the half-scale section for mid, at or above which a mask stops contributing to the evidence. In the tiled passes this fraction is also taken of the whole section, not of the tile. The mask itself remains available as an atom. Evidence is the per-pixel maximum of the sigmoid of the mask logits, accumulated before non-maximum suppression. In the tiled passes, masks touching an inner tile border are dropped from the labels, evidence is merged across tiles by maximum, and labels are painted from the largest mask down. Each tile, or each section in the other passes, is stretched to its own minimum and maximum over all channels before detection.

#### C.1.5 DATASET-DEPENDENT QUANTITIES

Tile width and stride follow the shorter in-plane side, and the minimum sizes used by `augment`, `atomtrack` and `filter` scale with the in-plane area by fixed rules and are not set by hand. Appendix C.2 gives the rules and their reference values. The minimum coarse-mask area used by `parent` is the exception. It is held at 20,000 native pixels on every dataset and is not scaled. Every other parameter in this appendix is identical across all datasets reported in this paper. The hysteresis thresholds of `gate` are derived from each volume, and the contrast score $z$ of `filter` is standardized within each volume.

### C.2 THE FROZEN CONFIGURATION

This appendix gives the quantities of the frozen implementation. Appendix C.1 states the rules that use them.

#### C.2.1 FOREGROUND GATE

Each per-view evidence volume is smoothed with a Gaussian of $\sigma = 1$ voxel, and fused evidence $C$ is the mean of the two strongest smoothed values at each voxel (`FUSE=top2`). The per-view gamma hooks are the identity. With `SELFCAL=on` an Otsu threshold (Otsu, 1979) is computed on $C$. Then $b$ is the median of the values at or below it and $h$ the median of the values above it, and the two hysteresis thresholds are

$$\tau_{\text{lo}} = b + 0.68\,(h - b), \qquad \tau_{\text{hi}} = b + 0.87\,(h - b).$$

Seeds above $\tau_{\text{hi}}$ are grown through 6-connected voxels above $\tau_{\text{lo}}$. Both thresholds are therefore properties of the volume being segmented, not constants.

#### C.2.2 TRACK ACCEPTANCE

The `filter` stage accepts or rejects tracks based on four conditions. Write $\ell$ for the number of distinct sections a track occupies, $a$ for the median per-section area, $\bar{C}$ for the voxel-weighted mean of the fused evidence over the track, $\tilde{C}$ and $C_{25}$ for the median and first quartile of its per-section means, and $z$ for its contrast score. The contrast of a section is the guide $g$, rescaled to its 1 to 99 percentile range, minus a background obtained by a Gaussian of $\sigma = 8$ on the section downsampled fourfold. The score $z$ standardizes the mean contrast of the track against all track-section means of the same volume. The conditions are checked in the order below, and a track is recorded with the first condition it fails.

1. **Artifact.** Reject when $a < 16$, or when $\ell = 1$ and $a < 64$, unless $\bar{C} \geq 0.70$ and $z \geq 1.5$.
2. **Persistence.** A track with $\ell \geq 3$ passes. A track with $\ell = 2$ is kept when $\bar{C} \geq 0.45$ or $z \geq 1.0$. A track with $\ell = 1$ is kept when $\bar{C} \geq 0.70$, $z \geq 1.5$ and $a \geq 64$.
3. **Evidence.** Reject when $\tilde{C} < 0.30$ and $C_{25} < 0.20$.
4. **Contrast.** A track with $z \geq 0.5$ passes. A track with $z < -0.5$ and $\tilde{C} < 0.60$ is rejected. Any other track with $z < 0.5$ requires $\ell \geq 3$.

Tracks recorded as rejected by persistence or contrast are restored when their orthogonal support reaches $0.35$. Tracks rejected by evidence are restored at $0.50$. Artifact rejections are not restored.

### C.2.3 SIZES DERIVED FROM THE IMAGE

With $A$ the in-plane area of a section and $A_{\text{ref}} = 2000 \times 2500$ pixels, each size threshold is $\max(\text{floor},\ \text{round}(v_{\text{ref}} A / A_{\text{ref}}))$.

| quantity | $v_{\text{ref}}$ | floor |
|---|---|---|
| large atom | 50,000 | 1,000 |
| minimum logged volume | 2,000 | 50 |
| minimum coarse mask | 20,000 | 400 |
| minimum piece | 10,000 | 200 |
| minimum component | 5,000 | 100 |
| minimum mid-scale mask | 5,000 | 100 |
| minimum mid-scale piece | 2,500 | 50 |
| minimum mid-scale component | 1,250 | 25 |

On the $256^3$ cubes every value sits at its floor, and the large-atom size is $1{,}000$ pixels. The `parent` stage always receives $20{,}000$ native pixels, because the driver passes the scaled value to `augment` only.

Tiling follows the shorter in-plane side $m$: tile $512$ when $m \geq 1024$ or $m = 512$, else $256$ when $m \geq 512$ or $m = 256$, else $128$ when $m \geq 256$ or $m = 128$, else the tile is $m$ itself. Stride is half the tile. The mid branch applies the same rule to the half-scale image. For the cubes this gives tile $256$ and stride $128$ at native scale, and $128$ and $64$ at mid scale, which is a single tile covering the whole section at both scales.

The coarse pass resizes a section so that its longer in-plane side is at most $1024$ pixels and never enlarges. It therefore does not resize the cubes at all.

### C.3 COMPARED METHODS

Table 6 lists the version, weights and changed settings of every compared method.

Table 6: Compared methods and their invocation. Each method was run at its published defaults except for the settings in the last column.

| Method | Version, weights | Changed from default |
|---|---|---|
| VISTA3D (He et al., 2025) | NV-Segment-CT MONAI bundle 0.6.0 (f9f5f51b), HF nvidia/NV-Segment-CT | mode: everything |
| SAMamba3D (Zhang et al., 2026) | commit ff92f22d, samamba3d_turbo.pth | num_classes=4, patch $96^3$ / stride $48^3$ |
| Cellpose-SAM (Pachitariu et al., 2025) | cellpose 4.2.1.1, cpsam | volumetric inference |
| µSAM (AIS) (Archit et al., 2025) | micro_sam 1.8.6, vit_b_lm + AIS decoder | AIS + merge |
| DEVA (Cheng et al., 2023) | stock DEVA commit 2aafa2f0, SAM ViT-H (sam_vit_h_4b8939) and DEVA-propagation | automatic mode, $XY$ sections as frames, `--max_num_objects -1`, `--size -1` |
| u-Segment3D (Zhou et al., 2025) | u-Segment3D 0.1.4, backbone cellpose 3.1.1.3 cyto2 | direct, 3 views xy/xz/yz, gradient_decay=0.1 |
| SAM 3 (Carion et al., 2026) | sam3 0.1.0, facebook/sam3 | video mode, $XY$ sections as frames, text prompt "Everything" on the first frame |

## D ADDITIONAL RESULTS

This appendix collects the per-plane scores, the figures of all methods and the supporting tables cited in the main text.

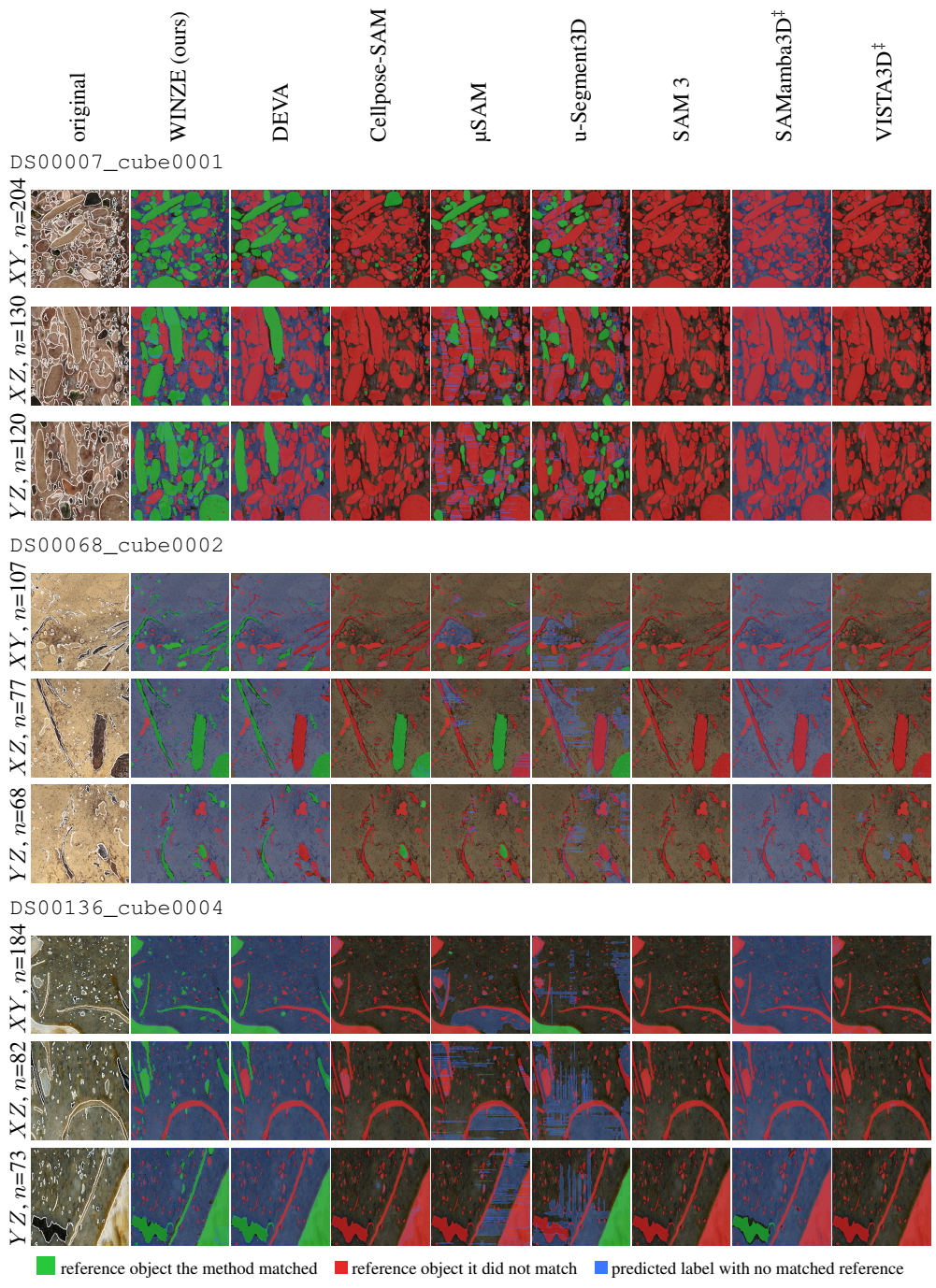



Figure 5: All results for eight methods on every scored plane of all three cubes. Rows are the three orientations of each cube, columns are the methods, and $n$ is the number of reference objects on that plane. The original column overlays the reference outlines in white. The scores for these planes are in Table 7. ‡ semantic output.

Table 7: Instance F1@0.5 on each scored plane. The three planes of a cube are one per orientation, and $n$ is the number of reference objects on that plane. The last column aggregates the three planes of that cube and repeats Table 2. Bold marks the single best method in a column. A column with a tie at the top carries no bold. ‡ semantic output.

| Method | $XY$ | $XZ$ | $YZ$ | 3 planes |
|---|---|---|---|---|
| `DS00007_cube0001` | | | | |
| | ($n$=204) | ($n$=130) | ($n$=120) | ($n$=454) |
| WINZE (ours) | **0.492** | **0.139** | 0.156 | **0.228** |
| DEVA | 0.159 | 0.059 | 0.072 | 0.111 |
| Cellpose-SAM | 0.159 | 0.026 | 0.029 | 0.088 |
| µSAM | 0.239 | 0.070 | 0.068 | 0.113 |
| u-Segment3D | 0.187 | 0.124 | **0.209** | 0.174 |
| SAM 3 | 0.000 | 0.000 | 0.000 | 0.000 |
| SAMamba3D‡ | 0.000 | 0.000 | 0.000 | 0.000 |
| VISTA3D‡ | 0.000 | 0.000 | 0.000 | 0.000 |
| `DS00068_cube0002` | | | | |
| | ($n$=107) | ($n$=77) | ($n$=68) | ($n$=252) |
| WINZE (ours) | **0.465** | **0.181** | **0.205** | **0.258** |
| DEVA | 0.081 | 0.093 | 0.073 | 0.082 |
| Cellpose-SAM | 0.000 | 0.049 | 0.055 | 0.030 |
| µSAM | 0.051 | 0.015 | 0.017 | 0.027 |
| u-Segment3D | 0.018 | 0.024 | 0.000 | 0.015 |
| SAM 3 | 0.000 | 0.000 | 0.000 | 0.000 |
| SAMamba3D‡ | 0.000 | 0.000 | 0.000 | 0.000 |
| VISTA3D‡ | 0.000 | 0.000 | 0.000 | 0.000 |
| `DS00136_cube0004` | | | | |
| | ($n$=184) | ($n$=82) | ($n$=73) | ($n$=339) |
| WINZE (ours) | **0.228** | **0.111** | **0.068** | **0.128** |
| DEVA | 0.080 | 0.061 | 0.044 | 0.067 |
| Cellpose-SAM | 0.000 | 0.000 | 0.000 | 0.000 |
| µSAM | 0.010 | 0.000 | 0.000 | 0.004 |
| u-Segment3D | 0.011 | 0.000 | 0.026 | 0.011 |
| SAM 3 | 0.000 | 0.000 | 0.000 | 0.000 |
| SAMamba3D‡ | 0.000 | 0.000 | 0.026 | 0.006 |
| VISTA3D‡ | 0.000 | 0.000 | 0.000 | 0.000 |

Table 8: Recall, precision and the number of predicted labels on the imaging plane and on the orthogonal planes, at IoU $\geq 0.5$. The imaging plane is the single scored $XY$ plane of a cube. The orthogonal columns pool its scored $XZ$ and $YZ$ planes. The last pair divides the number of predicted labels by the number of reference objects on the same planes. A value above one therefore means that the method returns more labels than there are objects. ‡ semantic output.

| | recall | | precision | | labels / reference | |
|---|---|---|---|---|---|---|
| Method | in-plane | orth. | in-plane | orth. | in-plane | orth. |
| `DS00007_cube0001` | | | | | | |
| WINZE (ours) | 0.461 | 0.376 | 0.528 | 0.092 | 0.87 | 4.08 |
| DEVA | 0.127 | 0.044 | 0.211 | 0.126 | 0.60 | 0.35 |
| Cellpose-SAM | 0.098 | 0.016 | 0.426 | 0.095 | 0.23 | 0.17 |
| µSAM | 0.147 | 0.100 | 0.638 | 0.053 | 0.23 | 1.90 |
| u-Segment3D | 0.162 | 0.152 | 0.221 | 0.179 | 0.73 | 0.85 |
| SAM 3 | 0.000 | 0.000 | 0.000 | 0.000 | 0.00 | 0.00 |
| SAMamba3D‡ | 0.000 | 0.000 | 0.000 | 0.000 | 0.01 | 0.02 |
| VISTA3D‡ | 0.000 | 0.000 | 0.000 | 0.000 | 0.04 | 0.06 |
| `DS00068_cube0002` | | | | | | |
| WINZE (ours) | 0.374 | 0.359 | 0.615 | 0.132 | 0.61 | 2.72 |
| DEVA | 0.047 | 0.048 | 0.294 | 0.304 | 0.16 | 0.16 |
| Cellpose-SAM | 0.000 | 0.028 | 0.000 | 0.400 | 0.06 | 0.07 |
| µSAM | 0.028 | 0.014 | 0.273 | 0.019 | 0.10 | 0.73 |
| u-Segment3D | 0.009 | 0.007 | 0.167 | 0.071 | 0.06 | 0.10 |
| SAM 3 | 0.000 | 0.000 | 0.000 | 0.000 | 0.00 | 0.00 |
| SAMamba3D‡ | 0.000 | 0.000 | 0.000 | 0.000 | 0.03 | 0.04 |
| VISTA3D‡ | 0.000 | 0.000 | 0.000 | 0.000 | 0.17 | 0.12 |
| `DS00136_cube0004` | | | | | | |
| WINZE (ours) | 0.185 | 0.232 | 0.298 | 0.056 | 0.62 | 4.14 |
| DEVA | 0.043 | 0.032 | 0.471 | 0.147 | 0.09 | 0.22 |
| Cellpose-SAM | 0.000 | 0.000 | 0.000 | 0.000 | 0.03 | 0.15 |
| µSAM | 0.005 | 0.000 | 0.143 | 0.000 | 0.04 | 1.11 |
| u-Segment3D | 0.005 | 0.006 | 0.250 | 0.056 | 0.02 | 0.12 |
| SAM 3 | 0.000 | 0.000 | 0.000 | 0.000 | 0.00 | 0.00 |
| SAMamba3D‡ | 0.000 | 0.006 | 0.000 | 0.167 | 0.02 | 0.04 |
| VISTA3D‡ | 0.000 | 0.000 | 0.000 | 0.000 | 0.11 | 0.07 |

Table 9: Predicted labels per public specimen, summarized over specimens. The best-IoU measure charges nothing for labels placed outside the reference object, and with one reference object per specimen only one label per specimen enters the score. A method emitting more labels has more candidates for the single match and is not charged for the rest. It still loses if it splits the reference object, because a fragment loses volume and its IoU falls. ‡ semantic output.

| Method | specimens | median labels | range |
|---|---|---|---|
| WINZE (ours) | 11 | 1,805 | 530–42,886 |
| DEVA | 11 | 36 | 15–125 |
| Cellpose-SAM | 11 | 85 | 4–1,616 |
| µSAM | 10 | 146.5 | 16–1,782 |
| u-Segment3D | 11 | 188 | 20–3,052 |
| SAM 3 | 11 | 0 | 0–0 |
| SAMamba3D‡ | 11 | 3 | 3–3 |
| VISTA3D‡ | 11 | 73 | 43–115 |




Figure 6: Results for the public specimen `DS00084 XCCa3MD` in 3D view. Each panel after the reference shows the single predicted label with the highest IoU against the reference object (Table 3). ‡ semantic output.




Figure 7: The annotated objects of the qualitative cube, grouped by whether WINZE recovered them. An object counts as recovered when a single WINZE label reaches IoU $\geq 0.5$ with it (40 of 53). All three panels show the annotated shapes in their natural color. The cube is rendered with equal edges.

Table 10: Foreground coverage against sectional F1@0.5. For each cube, *fg* is the fraction of a scored plane that a method labeled as foreground, taken as the median over the three planes, and *F1* is its sectional instance F1@0.5. A method that labels far less of a plane than the reference rarely reaches the matching stage, which is a failure to detect rather than a failure to separate. ‡ semantic output. SAMamba3D assigns every voxel to one of four material classes and carries almost no instance boundaries.

| | `DS00007_cube0001` | | `DS00068_cube0002` | | `DS00136_cube0004` | |
|---|---|---|---|---|---|---|
| Method | fg | F1 | fg | F1 | fg | F1 |
| WINZE (ours) | 0.955 | 0.228 | 0.985 | 0.258 | 0.984 | 0.128 |
| DEVA | 0.955 | 0.111 | 0.936 | 0.082 | 0.991 | 0.067 |
| Cellpose-SAM | 0.011 | 0.088 | 0.012 | 0.030 | 0.011 | 0.000 |
| µSAM | 0.244 | 0.113 | 0.148 | 0.027 | 0.204 | 0.004 |
| u-Segment3D | 0.300 | 0.174 | 0.267 | 0.015 | 0.414 | 0.011 |
| SAM 3 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SAMamba3D‡ | 1.000 | 0.000 | 1.000 | 0.000 | 1.000 | 0.006 |
| VISTA3D‡ | 0.007 | 0.000 | 0.013 | 0.000 | 0.002 | 0.000 |

Table 11: The 20-pixel secondary analysis in full. Objects smaller than 20 pixels in plane are removed from both the prediction and the reference before matching. The threshold is in pixels, not physical units, and therefore corresponds to a different physical size on each cube. The analysis was proposed after the first results were in and is labeled as secondary wherever it appears. In the ROI cubes an object is removed when its largest $XY$ cross-section is smaller than 20 pixels. 3 pl. is the three-plane scope. ‡ semantic output.

| | | DS00007_cube0001 | | DS00068_cube0002 | | DS00136_cube0004 | |
|---|---|---|---|---|---|---|---|
| Method | Analysis | 3 pl. | ROI | 3 pl. | ROI | 3 pl. | ROI |
| WINZE (ours) | all objects | 0.228 | 0.153 | 0.258 | 0.072 | 0.128 | 0.207 |
| | $\geq$ 20 px | 0.434 | 0.180 | 0.495 | 0.116 | 0.314 | 0.289 |
| DEVA | all objects | 0.111 | 0.000 | 0.082 | 0.071 | 0.067 | 0.040 |
| | $\geq$ 20 px | 0.121 | 0.000 | 0.136 | 0.118 | 0.129 | 0.071 |
| Cellpose-SAM | all objects | 0.088 | 0.000 | 0.030 | 0.074 | 0.000 | 0.000 |
| | $\geq$ 20 px | 0.086 | 0.000 | 0.053 | 0.125 | 0.000 | 0.000 |
| µSAM | all objects | 0.113 | 0.156 | 0.027 | 0.057 | 0.004 | 0.035 |
| | $\geq$ 20 px | 0.155 | 0.178 | 0.049 | 0.091 | 0.007 | 0.065 |
| u-Segment3D | all objects | 0.174 | 0.161 | 0.015 | 0.000 | 0.011 | 0.000 |
| | $\geq$ 20 px | 0.201 | 0.189 | 0.026 | 0.000 | 0.024 | 0.000 |
| SAM 3 | all objects | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | $\geq$ 20 px | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SAMamba3D‡ | all objects | 0.000 | 0.000 | 0.000 | 0.000 | 0.006 | 0.040 |
| | $\geq$ 20 px | 0.000 | 0.000 | 0.000 | 0.000 | 0.012 | 0.071 |
| VISTA3D‡ | all objects | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | $\geq$ 20 px | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| reference objects | all | 454 | 57 | 252 | 26 | 339 | 47 |
| | $\geq$ 20 px | 419 | 50 | 140 | 15 | 162 | 25 |

Table 12: Number of predicted labels in the bundle of each true positive when a bundle is allowed to match one reference object. A bundle with $k = 1$ is a single label, which is the rule of every primary score. The bundle size differs by method. A predicted label joins the bundle of the reference object that holds more than half of its voxels, and the bundles are then matched one to one as in Table 2. ALL pools the six scopes, ALL-planes the three plane scopes and ALL-roi the three ROI scopes. VISTA3D and SAM 3 have no true positive and are not listed. ‡ semantic output.

| Scope | Method | TP | median $k$ | max $k$ | $k = 1$ | $k \geq 5$ |
|---|---|---|---|---|---|---|
| ALL | WINZE (ours) | 397 | 1.0 | 21 | 81.4% | 5.3% |
| ALL | DEVA | 68 | 1.0 | 2 | 86.8% | 0.0% |
| ALL | Cellpose-SAM | 31 | 1.0 | 4 | 87.1% | 0.0% |
| ALL | µSAM | 94 | 1.0 | 17 | 60.6% | 13.8% |
| ALL | SAMamba3D‡ | 2 | 1.0 | 1 | 100.0% | 0.0% |
| ALL | u-Segment3D | 91 | 1.0 | 5 | 89.0% | 1.1% |
| ALL-planes | WINZE (ours) | 353 | 1.0 | 21 | 81.3% | 4.8% |
| ALL-planes | DEVA | 66 | 1.0 | 2 | 87.9% | 0.0% |
| ALL-planes | Cellpose-SAM | 30 | 1.0 | 4 | 86.7% | 0.0% |
| ALL-planes | µSAM | 76 | 1.0 | 17 | 63.2% | 13.2% |
| ALL-planes | SAMamba3D‡ | 1 | 1.0 | 1 | 100.0% | 0.0% |
| ALL-planes | u-Segment3D | 79 | 1.0 | 5 | 91.1% | 1.3% |
| ALL-roi | WINZE (ours) | 44 | 1.0 | 13 | 81.8% | 9.1% |
| ALL-roi | DEVA | 2 | 1.5 | 2 | 50.0% | 0.0% |
| ALL-roi | Cellpose-SAM | 1 | 1.0 | 1 | 100.0% | 0.0% |
| ALL-roi | µSAM | 18 | 1.5 | 11 | 50.0% | 16.7% |
| ALL-roi | SAMamba3D‡ | 1 | 1.0 | 1 | 100.0% | 0.0% |
| ALL-roi | u-Segment3D | 12 | 1.0 | 2 | 75.0% | 0.0% |

Table 13: clDice medians over reference objects, with the number of objects on which WINZE scores exactly zero. In four of the six scopes at least five of the seven compared methods have a median of 0.000 and none exceeds 0.011. The median carries no ordering there, and the main table therefore reports the mean. WINZE's own median on the `DS00136_cube0004` planes is 0.014, with 109 of its 339 objects scoring exactly zero, while its mean is 0.326. 3 pl. is the three-plane scope. ‡ semantic output.

| | `DS00007_cube0001` | | `DS00068_cube0002` | | `DS00136_cube0004` | |
|---|---|---|---|---|---|---|
| Method | 3 pl. | ROI | 3 pl. | ROI | 3 pl. | ROI |
| WINZE (ours) | 0.490 | 0.555 | 0.293 | 0.365 | 0.014 | 0.566 |
| DEVA | 0.142 | 0.044 | 0.000 | 0.001 | 0.000 | 0.000 |
| Cellpose-SAM | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| µSAM | 0.000 | 0.213 | 0.000 | 0.000 | 0.000 | 0.000 |
| u-Segment3D | 0.195 | 0.217 | 0.000 | 0.000 | 0.000 | 0.001 |
| SAM 3 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SAMamba3D‡ | 0.008 | 0.028 | 0.004 | 0.011 | 0.000 | 0.001 |
| VISTA3D‡ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| WINZE objects scoring zero | 37/454 | 1/57 | 66/252 | 5/26 | 109/339 | 6/47 |

Table 14: What the grouping stage of Section 3.4 does to the fragment count in the four cubes of this paper. The last column is the fraction of the whole volume held by the single largest instance after grouping. The accepted foreground is unchanged by this stage, and the reduction therefore redistributes identity rather than removing material. The column *fragments in* counts the tracklets rebuilt for the grouping (Section 3.4), not the labels of $\widehat{Y}_0$.

| Cube | fragments in | fragments out | reduction | largest share |
|---|---|---|---|---|
| `DS00007_cube0001` | 16,773 | 6,954 | 58.5% | 0.393 |
| `DS00068_cube0002` | 5,328 | 3,067 | 42.4% | 0.784 |
| `DS00136_cube0004` | 6,017 | 3,903 | 35.1% | 0.593 |
| `DS00173_cube0006` | 5,856 | 3,696 | 36.9% | 0.691 |

Table 15: Objects recovered at IoU $\geq 0.5$ in the qualitative cube, counted by stratum. The three groupings each partition the same 53 annotated objects, and the column headings give the number of objects in each stratum. The annotation is positive-only, and detection counts are therefore the only accuracy measure defined here. The strata were assigned by the annotator from the reference outlines alone. The shape stratum *thin* is a separate grouping from the category *thin bivalve shell* of Table 4. Appendix B.1 gives the provenance. The last two rows give the median over the objects in each stratum of their mean luminance and the median distance of that mean from the matrix luminance, which is 125.7 on the 0–255 scale (the median over all matrix voxels). Luminance is the mean of the three color channels. The matrix is every voxel that carries no reference label and therefore includes any object that was not annotated. ‡ semantic output.

| Method | color | | shape | | | size | | |
|---|---|---|---|---|---|---|---|---|
| | bright (22) | dark (31) | compact (29) | elongated (9) | thin (15) | large (16) | medium (19) | small (18) |
| WINZE (ours) | 14 | 26 | 25 | 6 | 9 | 14 | 15 | 11 |
| DEVA | 2 | 4 | 3 | 1 | 2 | 5 | 0 | 1 |
| Cellpose-SAM | 3 | 1 | 4 | 0 | 0 | 3 | 1 | 0 |
| µSAM | 1 | 2 | 3 | 0 | 0 | 1 | 2 | 0 |
| u-Segment3D | 1 | 1 | 1 | 1 | 0 | 2 | 0 | 0 |
| SAM 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| SAMamba3D‡ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| VISTA3D‡ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *median luminance* | 144.0 | 55.7 | 74.1 | 81.6 | 66.7 | 141.5 | 66.7 | 83.9 |
| *median distance from matrix* | 25.4 | 70.0 | 58.5 | 52.2 | 59.0 | 62.5 | 59.0 | 45.9 |

# E ADDITIONAL EXPERIMENTS

**Sectional masks linked by overlap alone.** To separate the quality of the sectional masks from the assembly across sections, we linked the masks that the SAM 2 checkpoint of WINZE generates automatically on each $XY$ section along $Z$ by overlap alone, with no other stage of WINZE. Table 16 gives the settings. On `DS00007_cube0001` and `DS00136_cube0004` these masks are precise in plane, and the scores collapse on the orthogonal planes. The cube `DS00068_cube0002` does not test the linking, for the reason given in Table 16. WINZE scores higher than the baseline in all four scopes of the other two cubes. We also classified the unmatched labels on the orthogonal planes by where they lie (Table 17). Most unmatched labels of the baseline lie inside a single reference object, which is the signature of one object split along $Z$. Most unmatched labels of WINZE lie largely outside the reference objects.

Table 16: Experiment on sectional detection linked along $Z$ without the assembly of WINZE. The same SAM 2 checkpoint as WINZE (`sam2.1_hiera_tiny`) ran automatic mask generation on the whole of each $XY$ section with 64 points per side. The mask thresholds were the defaults of the generator (pred_iou_thresh 0.8, stability_score_thresh 0.95), not those of WINZE (Appendix C.1.4). Masks of adjacent sections were joined along $Z$ by greedy one-to-one matching at IoU $> 0.5$, and no other stage of WINZE was used. Overlaps within a section were resolved in favor of the mask with the higher predicted score. One-to-one linking leaves the partition of each section unchanged, which means that the in-plane values measure the quality of the 2D masks. On `DS00068_cube0002` a single mask covering almost the whole section had the highest score on 218 of the 256 sections, and the overlap rule removed the other masks. The values on this cube therefore reflect that rule and not identity across sections. Linking at IoU $> 0.3$ or $> 0.7$ instead changes any value in the table by at most 0.015. All rows are scored with the scorer of Table 2.

| | `DS00007_cube0001` | | `DS00068_cube0002` | | `DS00136_cube0004` | |
|---|---|---|---|---|---|---|
| *(a) instance F1@0.5* | | | | | | |
| Method | 3 planes | ROI cube | 3 planes | ROI cube | 3 planes | ROI cube |
| SAM 2 per-section + IoU link | 0.130 | 0.088 | 0.005 | 0.022 | 0.094 | 0.075 |
| WINZE (ours) | 0.228 | 0.153 | 0.258 | 0.072 | 0.128 | 0.207 |
| u-Segment3D | 0.174 | 0.161 | 0.015 | 0.000 | 0.011 | 0.000 |
| *(b) sectional F1@0.5* | | | | | | |
| Method | in-plane | orth. | in-plane | orth. | in-plane | orth. |
| SAM 2 per-section + IoU link | 0.320 | 0.054 | 0.000 | 0.008 | 0.198 | 0.028 |
| WINZE (ours) | 0.492 | 0.148 | 0.465 | 0.193 | 0.228 | 0.090 |
| u-Segment3D | 0.187 | 0.165 | 0.018 | 0.013 | 0.011 | 0.012 |
| *(c) precision* | | | | | | |
| Method | in-plane | orth. | in-plane | orth. | in-plane | orth. |
| SAM 2 per-section + IoU link | 0.870 | 0.045 | 0.000 | 0.009 | 0.579 | 0.025 |
| WINZE (ours) | 0.528 | 0.092 | 0.615 | 0.132 | 0.298 | 0.056 |
| u-Segment3D | 0.221 | 0.179 | 0.167 | 0.071 | 0.250 | 0.056 |

Table 17: Composition of the unmatched predicted labels on the orthogonal planes. The planes are the scored $XZ$ and $YZ$ planes of the three annotated cubes, and the counts are summed over the three cubes before the fractions are taken. A predicted label is unmatched when the one-to-one assignment of Table 2 gives it no reference object at IoU $\geq 0.5$. An unmatched label lies inside one object when at least half of its area overlaps a single reference object. It lies across objects when no single object reaches half of its area but the reference objects together cover at least half of it. All remaining unmatched labels are counted as mostly outside reference objects, which includes labels that straddle an object boundary. Across area thresholds of 0.3, 0.5 and 0.7, the inside-one-object fraction ranges from 0.158 to 0.301 for WINZE and from 0.525 to 0.712 for SAM 2 with IoU linking. The SAM 2 row is the sectional detection with IoU linking of Table 16. SAM 3 returned no predicted label and is not listed. ‡ semantic output.

| | | fraction of unmatched labels | | |
|---|---|---|---|---|
| Method | unmatched | inside one object | across objects | mostly outside reference objects |
| WINZE (ours) | 1874 | 0.243 | 0.016 | 0.741 |
| SAM 2 per-section + IoU link | 667 | 0.661 | 0.021 | 0.318 |
| DEVA | 121 | 0.190 | 0.223 | 0.587 |
| Cellpose-SAM | 67 | 0.910 | 0.000 | 0.090 |
| µSAM | 727 | 0.631 | 0.026 | 0.343 |
| u-Segment3D | 204 | 0.397 | 0.098 | 0.505 |
| SAMamba3D‡ | 17 | 0.118 | 0.647 | 0.235 |
| VISTA3D‡ | 42 | 0.548 | 0.024 | 0.429 |

Table 18: Color against luminance. The in-plane and orthogonal columns are sectional F1@0.5 under the two inputs, and the gap is in-plane minus orthogonal. The columns headed *all* give the three-plane instance F1@0.5. The change in the gap is given as a difference, not as a ratio. The gap of WINZE widens on the development cube and narrows slightly on the two independent cubes. u-Segment3D, VISTA3D and SAMamba3D convert the input to luminance internally and are omitted. SAM 3 returned no object under either input and does not enter the counts in the text. The color columns come from the run made for this comparison. Their values in the *all* columns match Table 2 except for DEVA on `DS00007_cube0001`. Both conditions here come from that one run, which is what the comparison requires.

| | | color | | | grayscale | | | |
|---|---|---|---|---|---|---|---|---|
| Cube | Method | in-pl. | orth. | all | in-pl. | orth. | all | $\Delta$ gap |
| `DS00007_cube0001` | WINZE (ours) | 0.492 | 0.148 | 0.228 | 0.524 | 0.074 | 0.140 | +0.106 |
| `DS00007_cube0001` | Cellpose-SAM | 0.159 | 0.027 | 0.088 | 0.313 | 0.201 | 0.253 | −0.020 |
| `DS00007_cube0001` | µSAM | 0.239 | 0.069 | 0.113 | 0.252 | 0.061 | 0.110 | +0.020 |
| `DS00007_cube0001` | DEVA | 0.146 | 0.064 | 0.104 | 0.106 | 0.060 | 0.082 | −0.036 |
| `DS00007_cube0001` | SAM 3 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| `DS00068_cube0002` | WINZE (ours) | 0.465 | 0.193 | 0.258 | 0.450 | 0.184 | 0.249 | −0.007 |
| `DS00136_cube0004` | WINZE (ours) | 0.228 | 0.090 | 0.128 | 0.205 | 0.087 | 0.118 | −0.020 |

Table 19: Measured cost for one $256^3$ cube (`DS00007_cube0001`) under the frozen configuration. This is a single measurement, not a mean over repeats. It was taken on one RTX 6000 Ada with 48 GiB of device memory.

| Stage | Where | min | % |
|---|---|---|---|
| native | GPU detect | 8.71 | 25.9 |
| mid | GPU detect | 8.79 | 26.2 |
| cross | GPU detect | 4.61 | 13.7 |
| coarse | GPU detect | 8.74 | 26.0 |
| GPU detect | | 30.86 | 91.9 |
| gate | CPU post | 0.35 | 1.0 |
| xview | CPU post | 0.03 | 0.1 |
| smooth | CPU post | 1.86 | 5.5 |
| postxview | CPU post | 0.03 | 0.1 |
| augment | CPU post | 0.07 | 0.2 |
| atomtrack | CPU post | 0.02 | 0.1 |
| filter | CPU post | 0.08 | 0.2 |
| ortho | CPU post | 0.02 | 0.1 |
| parent | CPU post | 0.02 | 0.1 |
| grouping | CPU post | 0.14 | 0.4 |
| CPU post | | 2.64 | 7.9 |
| TOTAL | | 33.58 | 100.0 |

Table 20: Points per side (PPS) for the native detection pass. The asterisk marks the frozen configuration, and $n$ is the number of reference objects. This is a post-hoc ablation and not how the value was chosen. The parameter was set to 64 from 96 before the cubes scored here were annotated, based on a separate dataset that is not used in any reported evaluation and on a 2.26-fold speedup of native detection measured at that time. F1 is nearly flat over this range and no setting is best everywhere. The value 64 is best in three of the six scopes. Label counts grow by 47 to 74% from 32 to 64 but by only 6 to 9% from 64 to 96. Bold marks the best setting in each row. Timings of PPS 32 and 96 are one measurement per cube on one RTX 6000 Ada, averaged over the three cubes. The PPS 64 timings are the single `DS00007_cube0001` measurement of Table 19, taken on a separate day. PPS 96 takes about 30% more total time than PPS 64 and 63% more than PPS 32.

| Scope | $n$ | PPS 32 | PPS 64* | PPS 96 |
|---|---|---|---|---|
| `DS00007_cube0001` 3 planes | 454 | **0.241** | 0.228 | 0.215 |
| `DS00007_cube0001` ROI cube | 57 | 0.159 | 0.153 | **0.164** |
| `DS00068_cube0002` 3 planes | 252 | 0.230 | **0.258** | 0.247 |
| `DS00068_cube0002` ROI cube | 26 | 0.071 | **0.072** | 0.068 |
| `DS00136_cube0004` 3 planes | 339 | **0.136** | 0.128 | 0.135 |
| `DS00136_cube0004` ROI cube | 47 | 0.119 | **0.207** | 0.190 |
| *native detection (min)* | | 2.4 | 8.7 | 19.2 |
| *one cube, total (min)* | | 26.9 | 33.6 | 43.7 |

Table 21: Instance F1@0.5 of WINZE and every compared method on one public volume from each of six other domains and on the three annotated cubes of this paper. The settings are those of the main experiments and are listed in Table 6. Underlined cells mark the domain each method was built for. SAMamba3D was built for phase classification of pore-scale images, which none of these columns represents. The public ground truths annotate only some objects (two tracks in the video), and a method that segments everything is penalized for every unannotated object it returns. These F1 values are therefore lower bounds against partial references, not exhaustive scores. The last three columns repeat the three-plane scope of Table 2. The datasets are Ovules (Wolny et al., 2020), NucMM (Lin et al., 2021), ParticleSeg3D (Gotkowski et al., 2024), VerSe (Sekuboyina et al., 2021; Löffler et al., 2020; Liebl et al., 2021), MitoEM (Wei et al., 2020), and DAVIS 2017 `bmx-trees` (Pont-Tuset et al., 2017). A dash marks no result. [a] The script that produced the cross-domain VISTA3D condition was not kept, and the rebuilt recipe did not reproduce the stored output. [b] Out of GPU memory on our workstation. ‡ semantic output.

| | public volume, other domains | | | | | | natural-color tomography | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Plant confocal | µCT nuclei | µCT particles | Clinical CT | EM mitoch. | Video | DS00007 | DS00068 | DS00136 |
| WINZE (ours) | 0.099 | 0.127 | 0.238 | 0.000 | 0.059 | 0.000 | <u>0.228</u> | <u>0.258</u> | <u>0.128</u> |
| DEVA | 0.008 | 0.039 | 0.155 | 0.000 | –[b] | <u>0.000</u> | 0.111 | 0.082 | 0.067 |
| Cellpose-SAM | <u>0.584</u> | 0.736 | 0.002 | 0.014 | 0.002 | 0.000 | 0.088 | 0.030 | 0.000 |
| µSAM | <u>0.270</u> | 0.000 | 0.442 | 0.000 | 0.000 | 0.000 | 0.113 | 0.027 | 0.004 |
| u-Segment3D | <u>0.233</u> | 0.000 | 0.195 | 0.018 | 0.026 | 0.000 | 0.174 | 0.015 | 0.011 |
| SAM 3 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | <u>0.000</u> | 0.000 | 0.000 | 0.000 |
| SAMamba3D‡ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.006 |
| VISTA3D‡ | 0.000 | 0.000 | 0.000 | <u>0.894</u> | 0.000 | –[a] | 0.000 | 0.000 | 0.000 |

## F Computing environment

WINZE and the seven compared methods were run on a single workstation, a Supermicro AS -4125GS-TNRT server. This unit is equipped with two CPUs (AMD EPYC 9374F, 32 cores, base clock 3.85 GHz) and three GPUs (NVIDIA RTX 6000 Ada, 48 GiB each) mounted on a motherboard (Supermicro Super H13DSG-O-CPU), and 1.5 TB of RAM in 24 modules of 64 GB DDR5 (Micron MTC40F2046S1RC48BA1). WINZE, Cellpose-SAM, µSAM, VISTA3D, u-Segment3D and SAM 3 were run natively on Windows, and SAMamba3D and DEVA under WSL2 (Ubuntu 22.04.2 LTS).