Foveate: Recursive Zoom for Training-Free, In-Context Instance Segmentation

Robert Andreas Leist[1], Thiago Santos Gouvea[1,2], and Daniel Sonntag[1,2]

[1] German Research Center for Artificial Intelligence (DFKI), Interactive Machine Learning, Oldenburg & Saarbrücken, Germany
{robert.leist, thiago.gouvea, daniel.sonntag}@dfki.de
[2] Carl von Ossietzky University of Oldenburg, Applied AI, Oldenburg, Germany

Abstract. Research such as marine ecology relies heavily on the analysis of corals and their morphologies from images, yet segmentation of instances remains a hard computer vision task: even foundation models such as SAM 3 typically require extensive annotation and retraining to perform well on them. We present Foveate, a training-free method that segments instances recursively leveraging frozen DINOv3 (24) embeddings, given as few as one exemplar mask of a target instance. Foveate mimics the human fovea by recursively redirecting the vision encoder's attention to candidate regions: starting from the whole image, it gates patches by their similarity to an exemplar bank, proposes tighter crops from the connected components of the foreground, and re-embeds each crop, iterating until every instance is separated and re-identified against the exemplars at its own scale. No fine-tuning, decoder, class labels, or full annotations are required. Because the exemplar defines the target *scale and granularity*, a single prompt discovers *every* instance of the prompted concept, both within an image and across images. As a proof of concept, we evaluate on two custom marine benchmarks — coral and polyp segmentation — and approach them training-free with roughly 40× fewer parameters than a promptable foundation model. Polyps remain hard for all methods, but an oracle-foreground analysis shows that most of the gap is recoverable with a stronger foreground step, isolating it as the key bottleneck. We share our design, early results, and open questions to gather feedback.

Keywords: Instance segmentation · Training-free · In-context · Self-supervised features

1 Introduction

Corals are a major driver of marine biodiversity and the effects of climate change on their survival are still being actively researched (9). Substantial effort therefore goes into studying their survival and growth factors. To this end, corals are grown in controlled laboratory environments and their morphologies are analysed from photographs, where each instance — a coral fragment or an individual polyp — is manually labelled and classified. This process is time-consuming and

tedious, so researchers frequently resort to coarse but fast annotation (e.g. single points instead of full outlines). As a result, the quality of downstream insights drops and important findings may be missed.

Prompted segmentation models such as the SAM family (4; 12; 23) aim to mitigate this by enabling fast yet rich instance annotation: users provide simple prompts such as points or bounding boxes, and the model turns them into an outline of the target instance. However, these models often generalize poorly to unseen domains. This is commonly addressed by retraining SAM on large domain-specific datasets (1; 13; 15; 17; 18; 22). While this demonstrates the models' wide applicability, it circles back to the original problem: the need for large amounts of annotated ground truth.

Self-supervised models offer a way out: they train on unannotated inputs alone and can be adapted to a new domain without any ground truth, albeit only for self-supervised objectives such as image reconstruction. The DINO family (5; 21; 24) produces semantically aligned, dense features that are commonly used for linear probing. Crucially, these features can also be grouped by object identity using cosine similarity (e.g. via agglomerative clustering) without any learning. This raises the question of how far instance-level segmentation can be pushed *without any training at all*.

Training-free, in-context approaches exist for semantic segmentation. For example, INSID3 (8) gates patches by cosine similarity to a reference image-mask pair, but it operates at one fixed resolution: the encoder spends the same patch budget on the whole image regardless of where the instances are. Small or densely packed instances, as in our coral and polyp benchmarks, fall below this effective patch resolution and merge into a single foreground blob. Additionally, semantic approaches do not yield the necessary individuation of objects (i.e. instances) that ecological analysis requires.

We take inspiration from the human fovea: rather than processing the entire visual field at uniform resolution, attention is repeatedly redirected to fixate, refine, and individuate. **Foveate** realizes this as a recursive cascade (Fig. 1). Starting from the full image, it (i) finds concept-similar regions, (ii) proposes tighter child crops, and (iii) re-embeds each crop at higher effective resolution, recursing until the similarity to the exemplars degrades.

Contributions. Our contributions are: (1) **Foveate**, a training-free, in-context, recursive cascade for instance segmentation on frozen DINOv3 features addressing a series of limitations of related work; (2) a proof of concept on two custom marine benchmarks for coral and polyp instance segmentation demonstrating the ecological use case; and (3) an analysis of where the recursive zoom helps and where it breaks down.

Code and data. We publish our code on Github: `https://github.com/robertleist/foveate/releases/tag/cv4e%40eccv26`. The data is owned and will be published by the Helmholtz Institute for Functional, Marine Biodiversity (HIFMB[3]).

[3] `https://hifmb.de`, Accessed: 08/2026

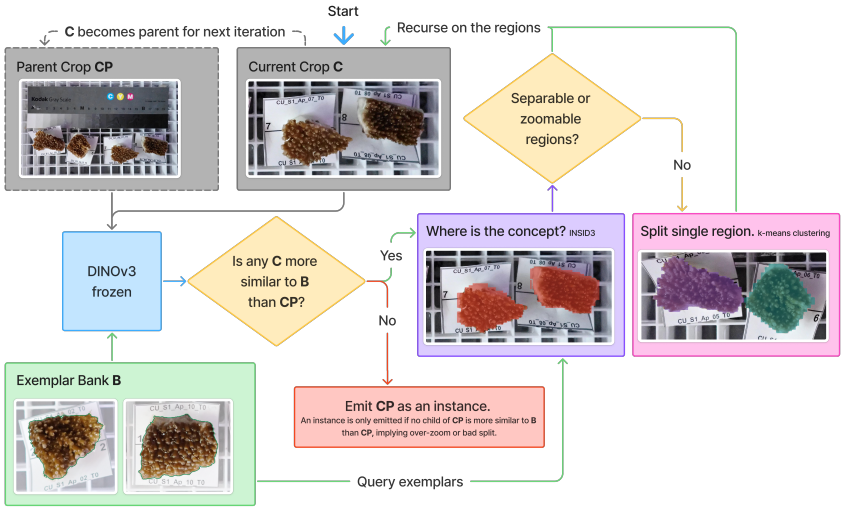


Fig. 1: One iteration of Foveate (simplified). The current crop C and the parent crop CP it came from are embedded by the frozen DINOv3 encoder and scored against the exemplar bank B. If no child of CP matches B more strongly than CP itself, the zoom has gone too far and CP is emitted as an instance. Otherwise `Where` (INSID3) locates the concept inside C, and the connected components of that foreground are either recursed into or, when a single region cannot be separated further, partitioned by $k = 2$-means (`Split`) before recursing. The figure omits the size floor ρ and the final NMS; see Algorithm 1 for the exact control flow.

2 Related Work

We review the ecological motivation for coral analysis and three lines of work that Foveate builds on: frozen self-supervised features, prompted segmentation for annotation, and training-free in-context segmentation.

2.1 Ecological importance of corals

Coral reefs cover a fraction of the ocean floor yet host a disproportionate share of marine biodiversity, and they are among the ecosystems most threatened by ocean warming and acidification (6; 20; 26). Understanding how corals grow and survive under changing conditions requires quantifying their morphology at the level of individual fragments and polyps, typically from large collections of laboratory photographs. Manual instance annotation of these images is a major bottleneck: it is slow, requires trained ecologists, and scales poorly with the number of instances per image (hundreds of polyps on a single fragment). This motivates annotation methods that generalize to the coral domain without per-domain retraining or large annotated corpora.

Assigning a class label to every pixel is called *semantic segmentation*, whereas detecting and outlining each object separately is *instance segmentation* (11). Crucially, coral analysis needs *instance-level* masks, not merely semantic ones. The quantities of interest are computed per object: fragment and polyp counts, per-fragment average color for bleaching assessment, and per-instance size and morphology (e.g., two adjacent corals of the same species must be told apart). A semantic mask that labels every coral pixel as "coral" merges touching fragments into one region and makes these measurements ill-defined: one cannot average color over the correct area, nor count what has not been separated. This distinction is often overlooked because the training-free and in-context methods that are practical under a no-ground-truth regime (Sec. 2.4) predominantly target semantic or concept segmentation and stop at exactly this granularity. Instance segmentation, though harder, is what the downstream ecology actually consumes.

2.2 Frozen self-supervised features for dense tasks

Unsupervised learning has been a pillar of machine learning from the start, initially in the form of clustering methods that group data points into classes (e.g. k-means or agglomerative clustering). Self-supervision extends this idea by letting a model generate its own supervisory signal (25). Such methods are especially popular in computer vision, where ground-truth labels are scarce and annotation is costly. Common approaches include contrastive learning and teacher-student distillation (25); in the latter, a student network learns to match a teacher's output under different augmentations of the image (5; 21; 24) or of its latent features (2; 3).

The DINO family follows this teacher-student paradigm (5; 21; 24). Its training regime repeatedly presents multiple views of an image (global and local crops) that the model must reconcile, yielding highly semantic, dense features. The most recent iteration, DINOv3 (24), adds Gram anchoring to stabilize dense features across long training runs, further improving their semantic quality. Foveate operates directly on these features, zooming into highly similar regions — a natural fit, since the backbone is already trained to be consistent across views and scales.

2.3 Prompted segmentation for annotation workflows

Annotation workflows trade off efficiency against accuracy: when a workflow is slow, annotators resort to coarser labels (e.g. point annotations instead of full outlines), sacrificing accuracy for speed. Prompted segmentation aims to make annotation both efficient and accurate. It refers to segmenting one or more objects from simple prompts such as points, boxes, or nouns (e.g. "cat"), covering both prompted *object* and prompted *concept* segmentation.

The SAM family (4; 12; 23) has shaped this field, providing foundation models with robust performance in the general domain. While SAM 1–2 perform prompted object segmentation only, SAM 3 (4) extends this to segment *every* instance of a concept defined by exemplar bounding boxes or a noun prompt,

further reducing annotation effort. SAM natively operates within a single image (intra-image): the exemplars must lie on the same image in which instances are to be found. Cross-image prediction, where the exemplars come from a different image than the target, is not natively possible with SAM 3, a real limitation of the foundation model. Moreover, SAM models often struggle to generalize to unseen domains, prompting a range of studies that retrain them for specific domains such as cellular, microscopy, marine, or medical segmentation (1; 13; 15; 17; 18; 22). While this again demonstrates the models' wide applicability, it presupposes large annotated datasets that are typically unavailable in research domains.

2.4 Training-free, in-context methods for segmentation

Training-free or zero-shot approaches use context to guide segmentation, eliminating the need to train a model for each unseen domain — inspired by the success of large language models that condition on in-context examples rather than being retrained. Context can be any additional input to the model that could inform segmentation. For example, the aforementioned prompted segmentation models are also in-context methods, where the prompts constitute the context.

SANSA reuses SAM 2's memory bank — originally used to track instances through videos — to transfer reference segmentations to new targets, effectively implementing cross-image prompted object segmentation (7). Similarly, Matcher finds corresponding anchor points via the cosine similarity of DINOv2 features between reference and target images and uses them as prompts for SAM 2 (16). Jo *et al.* (10) reduce the number of SAM 3 prompts per image via Chain-of-Prompting to just one per class, improving over the SAM 3 baseline. Other methods, such as SegIC, pretrain a mask decoder on top of a frozen DINOv2 backbone to consume the context as input (19). Most relevant to us, Cuttano *et al.* (8) introduce INSID3, a training-free, in-context framework based solely on DINO features: target embeddings are clustered, candidate clusters are selected by cross-similarity to a reference prototype, and clusters are then re-merged by self-similarity to a seed cluster. Two parameters guide this process: τ_{fg} controls the granularity of candidate clusters, while the aggregate threshold controls which clusters to merge with the seed. We directly incorporate INSID3 into our own algorithm as a foreground extraction method (see Sec. 3).

Several limitations recur across the previous methods. First, many rely on SAM's decoder and inherit its failure modes: when SAM fails, so do they, with only costly annotation and re-training as the remedy. Second, most fix their context to a pre-defined, single reference image-mask pair and offer no principled way to select it for real-world deployment. Third, they mostly operate either intra- or cross-image, solving only part of the problem. Fourth, a fixed input resolution caps the smallest resolvable instance, so small or densely packed objects are lost. Finally, none of them individuates instances of the same concept — the instance-level granularity that, as argued above, coral analysis requires.

3 Method

To overcome these limitations we propose Foveate, a training-free instance segmentation method that leverages a frozen DINOv3 backbone to recursively segment object instances from as few as one exemplar instance taken either from the target image itself (intra-image) or from a different one (cross-image). Foveate is a breadth-first recursive zoom (Algorithm 1) taking an image and exemplars as inputs and outputting all other instances of the same concept. Starting from the whole image, it maintains a frontier of crops. Each crop is embedded once and answers two questions: *where is the concept?* and, once the crop has converged, *is this a single instance?* Example instances, called *exemplars*, guide both questions at every step.

Exemplar bank. We first embed the exemplars once into a *bank* B. For each exemplar, we crop it from its original image and store the crop's patch embedding and CLS token in the bank. Throughout, we use a frozen DINOv3 ViT-S/16 encoder with $\sim$21.6M parameters.

Where is the concept? For the current crop, we embed it and extract the concept foreground (`Where`). We adopt the training-free INSID3 method (8), run against the top-1 bank exemplar whose CLS token is most similar to the current crop, where similarity is given by the cosine similarity. Let $\mathrm{cls}(\cdot)$ be the L2-normalized CLS token and c be our current crop and b an exemplar taken from the bank. We compute cosine similarity as

$$\mathrm{cs}(c, b) \;=\; \big\langle \mathrm{cls}(c),\, \mathrm{cls}(b) \big\rangle, \tag{1}$$

and refer to it as crop similarity in the following.

Extract and zoom. We extract instance candidates from the INSID3 foreground F_c returned by `Where` by taking its connected components (CC; `Extract`). This yields one of three cases: (i) multiple foreground components, so we recurse into one tighter crop per component; (ii) a single component strictly tighter than the crop, so we zoom into it; or (iii) a converged crop — a single component no tighter than the crop itself. Cases (i) and (ii) are intuitive zooms. Case (iii) arises when a converged crop still contains multiple instances that CC cannot separate (e.g. dense crowds); we then attempt to split the masked features into two via $k = 2$-means clustering (`Split`). `Split` therefore targets *instance* separation, not foreground/background separation: it is the only mechanism that can individuate touching instances once they have merged into a single component.

Stopping. The *only* stopping signals are re-identification and a size floor. For re-identification, a child must match the exemplar bank at least as strongly as the crop it came from; for the size floor ρ (`min_crop`), a crop at or below ρ is emitted without further splitting. There is no depth cap and no learned localization head. We define the re-identification score as

$$g(c) \;=\; \tfrac{1}{|B|} \sum_{b \in B} \mathrm{cs}(c, b), \tag{2}$$

Algorithm 1 Foveate cascade for one concept.

Require: image I, exemplar bank B, INSID3 thresholds τ_{fg} and AggT, crop similarity floor τ_C, size floor ρ

frontier $\leftarrow \{I\}$; leaves $\leftarrow \emptyset$
while frontier $\neq \emptyset$ **do**
  pop crop c; embed patches and cls(c)
  $F_c \leftarrow \textsc{Where}(c, B)$ ▷ INSID3
  $K \leftarrow \textsc{Extract}(c, B)$ ▷ Connected components
  if size(c) $\leq \rho$ **then** emit each component of K to leaves
  else if $|K| \geq 2$ **then** push one tighter crop per component
  else if the single component is strictly tighter than c **then** push it
  else ▷ converged: split and re-identify
    $\{c_1, c_2\} \leftarrow \textsc{Split}(c)$ ▷ k=2-means
    if $\max_i g(c_i) > g(c)$ **then** push each c_i with $g(c_i) \geq \tau_C$
    else emit c to leaves
    end if
  end if
end while
return NMS(leaves)

the mean crop similarity of a crop c to exemplars in the bank. We show in Sec. 4.2 that c contains a single instance if no zoom can improve $g(\cdot)$. Finally, we deduplicate overlapping leaves with non-maximum suppression (NMS).

4 Experiments

4.1 Datasets

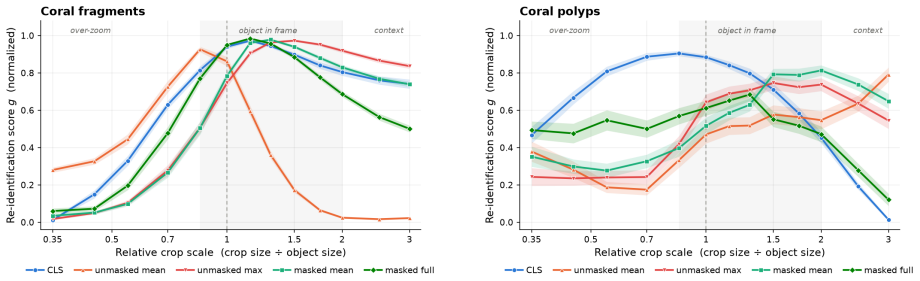
We use two custom instance-segmentation datasets of laboratory coral images provided by HIFMB. Both datasets share the same images and differ only in the target class and the number of fully annotated images.

The *corals* dataset consists of 199 coral-fragment instances annotated across 55 images. Marine ecologists use the average color of each coral to quantify bleaching under different conditions; here, point annotations do not suffice, and semantic segmentation fails whenever an image contains fragments from different corals.

The *polyps* dataset targets a finer coral morphology. Polyps sit densely on coral fragments and appear in large numbers, so this dataset contains 450 instances across only four images. For our purpose, this large instance count is sufficient to assess an in-context instance-segmentation method.

Both datasets were annotated with an open-source annotation tool IQUANA[4] by one of the authors, following training by marine ecologists.

[4] `https://github.com/Iquana-tool`, Accessed: 06/2026



(a) Coral fragments **(b)** Coral polyps

Fig. 2: Stopping signal $g(\cdot)$ against relative crop scale for five score variants, averaged over held-out instances (mean $\pm$ SEM, $n = 40$). A good signal rises as the object comes into frame and falls on over-zoom.

4.2 Stopping Signal Validation

Before running the instance-segmentation experiments, we validate the stopping signal $g(\cdot)$ in isolation. For a range of held-out target instances we generate a series of square crops centred on the instance, spanning from over-zoom (the crop lies *inside* the object) through the tight bounding box to zoomed-out context, and score each crop with $g(\cdot)$. Figure 2 plots the resulting curves.

We evaluate the stopping signal $g(\cdot)$ as defined in equation 2, as well as four other variants. (i) **CLS** is the crop similarity introduced above, i.e. the cosine similarity between the crop and exemplar `[CLS]` tokens. (ii) **unmasked mean** and (iii) **unmasked max** take the mean, respectively the maximum, cosine similarity between *all* crop patch features and the exemplar prototype. (iv) **masked mean** and (v) **masked full** restrict the comparison to the crop's foreground patches, pooling them to their mean prototype or matching them patch-to-patch to the exemplar patches. Foreground masks are produced by INSID3 (τ_{fg}=0.3, AggT=0.4).

Figure 2 shows that CLS yields a stable stopping signal on both datasets: it rises steadily as the crop tightens around the object, peaks at the true-instance scale, and falls under over-zoom. On coral fragments – large, sparsely distributed instances – all five scores behave well. On the densely packed polyps, however, the patch-based scores become unreliable: unmasked mean, unmasked max and masked mean peak too early, at zoomed-out crops (relative crop scale $s \approx 1.5$–2) that still contain several polyps, and would therefore halt Foveate on a coarse region before it isolates a single instance. The masked scores are additionally bounded by the quality of the INSID3 masks: given ground-truth masks they recover a well-centred peak (e.g., masked mean shifts from s=1.75 to s=1.07), whereas CLS and the unmasked scores do not use the masks and are unchanged. This foreshadows the suboptimal segmentation quality of INSID3 analysed in Sec. 4.7.

4.3 Protocol

We evaluate instance segmentation on the two custom marine datasets. We report mask *Average Precision* (AP) following the COCO protocol (14), i.e. the mean AP over IoU thresholds from 0.5 to 0.95. We distinguish *intra-image* AP (iAP), where the exemplar instances and the evaluated instances lie on the same image, from *cross-image* AP (cAP), where the exemplars come from a different image; note that iAP re-scores the exemplar instances themselves to ensure instance re-discovery. We further report the average of the two (AP) weighted by the amount of images in the intra- and cross-image set, together with inference throughput and peak GPU memory. We consider two exemplar regimes: a *multi-exemplar* setting, using 30% of instances as exemplars drawn from ten images for corals and one image for polyps, and a *single-exemplar* setting, using just one exemplar from one image for both datasets. Hyperparameters are tuned to maximize iAP on the intra-image set, replicating a scenario where experts would try the methods on a few images with exemplars before running them cross-image. Hence, cAP is not confounded with the hyperparameter selection.

4.4 Baselines & Implementation

We compare against two baselines that isolate the two pillars of our method: (i) *INSID3+CC*, a single INSID3 foreground pass over the whole image, split into instances by connected components — i.e. Foveate without recursion; and (ii) SAM 3 (4), a promptable foundation model, prompted two ways: with a noun prompt, and — for cross-image prediction — with the reference image and exemplar boxes concatenated to the target, following (8). Each method's hyperparameters are tuned for best results in the multi-exemplar case: INSID3's foreground threshold τ_{fg} and aggregation threshold for INSID3+CC and Foveate, and the confidence threshold t for SAM 3 (Tab. 4). For Foveate, we additionally ablate the split mechanism, comparing $k = 2$-means clustering against no split (accepting a converged crop directly). We omit INSID3's post-processing pipeline and simply use bilinear upsampling to fit masks to the original resolution. For the single-exemplar case, we reuse the best hyperparameters found in the multi-exemplar tuning.

Finally, we evaluate against an oracle foreground extractor to isolate its value. Oracle-Foveate replaces only the Where step: at each crop we take the patch-resolution, ground-truth foreground instead of the INSID3 mask, leaving the cascade, split, and re-identification unchanged. It is an upper bound on foreground quality, not a deployable configuration. Oracle+CC extracts connected components from the patch-resolution, ground-truth foreground mask.

We implement everything in PyTorch for tensor operations and evaluation, and use Hugging Face for the DINOv3 ViT-S/16 model. All experiments run on a single Nvidia RTX PRO 6000 GPU.

Table 1: Intra-/cross-image AP (i-/cAP) and average AP (AP) on the multi-exemplar marine datasets. All methods are training-free; Foveate and INSID3+CC share one frozen DINOv3 ViT-S/16 encoder. Models had 30% of instances available as exemplars from ten and one image for the corals and polyps set, respectively. Hyperparameters were swept per method. Oracle-Foveate replaces the INSID3 foreground with the ground-truth foreground, and Oracle+CC extracts connected components directly from it; both are upper bounds, not deployable methods. Best per column among the deployable (non-oracle) methods in **bold**.

		Corals			Polyps		
Method	Params	iAP	cAP	AP	iAP	cAP	AP
SAM 3 (4)	∼0.9B						
text prompt		**100**	26.9	40.2	8.2	0	2.1
image concat		**100**	**95.0**	**95.9**	8.2	3.5	4.6
INSID3+CC (single-pass)	21.6M	59.5	64.0	63.2	0	0	0
Foveate (ours)	21.6M						
w/ k-means split		62.4	44.8	48.0	3.3	0	0.8
w/o split		69.0	63.2	64.3	3.3	0	0.8
Oracle+CC (single-pass)	0	89.5	86.7	87.2	0.5	0.3	0.4
Oracle-Foveate	21.6M						
w/ k-means split		81.1	72.6	74.1	**23.9**	6.7	11.0
w/o split		84.8	82.8	83.1	**23.9**	**8.1**	**12.1**

4.5 Instance Segmentation

Tab. 1 reports the multi-exemplar results. The corals dataset is largely solved by SAM 3 with the image-concat workaround, whereas polyps are a very hard task on which all methods segment unreliably. Native SAM 3 with a text prompt is the weakest method on corals, underscoring the need for visual exemplars. INSID3+CC achieves reasonable AP on corals but fails to segment polyps entirely, which we attribute to polyps falling below the effective patch resolution of DINOv3 ViT-S/16. Foveate without splitting outperforms the INSID3+CC baseline, demonstrating the value of the recursive zoom — most clearly on polyps, where INSID3+CC finds nothing at all. Splitting with $k = 2$-means clustering, however, degrades performance substantially. Inspection shows that it does not separate clumps into distinct instances but instead splits them into an inner and an outer region: the outer region is roughly the same size and crop similarity as the parent, so all children survive even when their mean crop similarity worsens. We therefore conclude that $k = 2$-means is not an appropriate split here and needs to be refined (see Sec. 4.7). Finally, replacing INSID3 with an oracle foreground extractor improves AP substantially — from 64.3 to 83.1 on corals and from 0.8 to 12.1 on polyps, the latter surpassing SAM 3 (4.6) — confirming that the foreground (*Where*) step, not the zoom-and-stop idea, is a bottleneck; in annotation terms, a stronger *Where* would recover most of the missed instances. Even with the oracle, however, the $k = 2$-means split still trails no split (74.1

Table 2: Capability comparison. ✓: natively supported; (✓): partial or via a workaround; ✗: not supported. *X-img*: prompt with exemplars from a different image than the target; ∞ *inst.*: no cap on the number of predicted instances; *Multi-scale*: resolves sub-patch-scale instances by re-embedding; *Individuates*: separates touching instances of the same concept.

Method	Params	X-img	∞ inst.	Multi-scale	Individuates
SAM 3 (4)	∼0.9B	(✓)	✗	✗	✓
INSID3+CC (single-pass)	21.6M	✓	✓	✗	(✓)
Foveate (ours)	21.6M	✓	✓	✓	✓

Table 3: Intra-/cross-image AP (i-/cAP) and average AP (AP) on the marine datasets in the *single-exemplar* setting. All methods are training-free; Foveate and INSID3+CC share one frozen DINOv3 ViT-S/16 encoder. Each method uses a single exemplar from one image and the hyperparameter configuration tuned in the multi-exemplar case. Best per column in **bold**.

		Corals			Polyps		
Method	Params	iAP	cAP	AP	iAP	cAP	AP
SAM 3 (4) w/ image concat	∼0.9B	**100**	**94.7**	**94.8**	**8.2**	**2.4**	**3.8**
INSID3+CC (single-pass)	21.6M	80.0	63.1	64.3	0	0	0
Foveate (ours) w/o split	21.6M	50.3	43.2	43.3	1.6	0	0.4

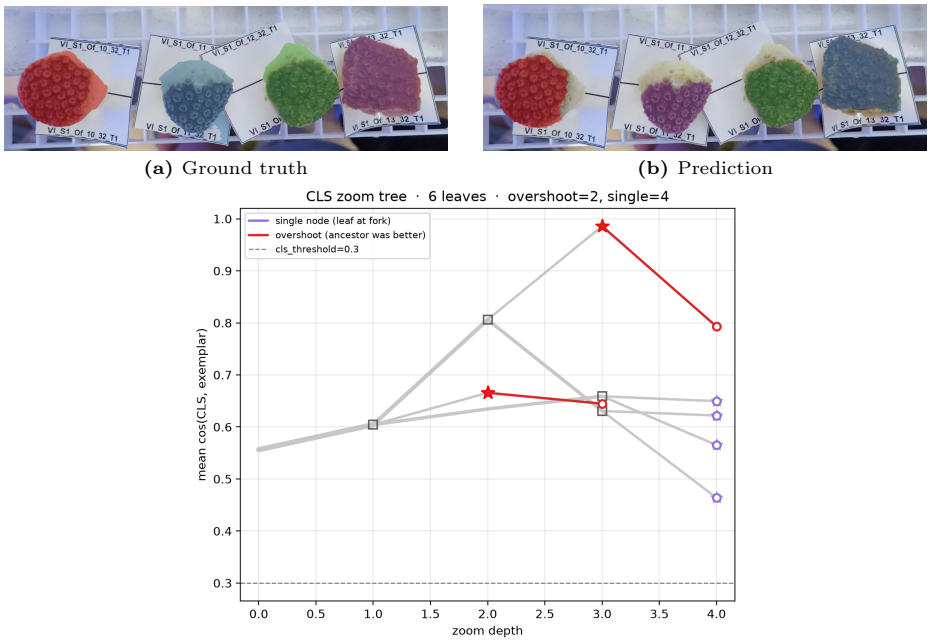
vs. 83.1 on corals; 11.0 vs. 12.1 on polyps), underscoring that a proper split function is needed as well. Moreover, comparisons to Oracle+CC demonstrate the effectiveness of the zoom for the polyps (0.4 vs. 12.1 AP). However, Oracle+CC outperforms Oracle-Foveate in the corals case highlighting that not just the *Where* step must be improved.

Raw AP, however, tells only part of the story. Tab. 2 contrasts the methods along capabilities that matter for deployment and explains why a lower-AP method can still be preferable: SAM 3 cannot prompt across images natively (requiring the image-concat workaround) and caps the number of predicted instances, whereas the single-pass INSID3+CC cannot resolve sub-patch instances and only partially individuates touching objects. Foveate is the only method that satisfies all four properties.

Tab. 3 reports the single-exemplar results. The overall picture matches the multi-exemplar case, with one notable exception: INSID3+CC now outperforms Foveate on corals. We attribute this to Foveate's scale-dependent oversegmentation, which we examine qualitatively in Sec. 4.7.

4.6 Efficiency

We report throughput (images/s) and peak GPU memory (MB) for the best configuration of each method. On corals, INSID3+CC and Foveate use only



(a) Ground truth **(b)** Prediction

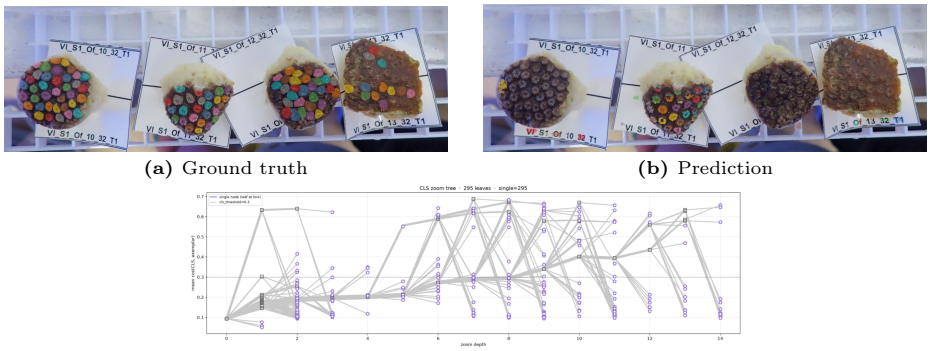
(c) Cascade

Fig. 3: One intra-image sample from the corals set. a) shows the ground truth annotations, while b) shows the predictions. c) shows the cascade as a connected graph, where each node represents one crop at a certain depth on the x axis and its crop similarity on the y axis.

600 MB versus 7373 MB for SAM 3; despite its recursion, Foveate still runs faster than SAM 3 (0.99 vs. 0.84 images/s), with the single-pass INSID3+CC fastest at 4.3 images/s. On polyps, throughput drops for all methods (1.3 images/s for INSID3+CC, ≈ 0.05 for SAM 3 and Foveate), but the memory gap widens sharply: SAM 3 peaks at 67,759 MB, whereas Foveate and INSID3+CC stay at 1333 and 1769 MB. Foveate thus attains competitive coverage on the hardest setting at a fraction of a foundation model's memory.

4.7 Qualitative insights

To validate our approach qualitatively, we show examples from the corals and polyps sets in Fig. 3 and Fig. 4, respectively. On corals, Foveate correctly localizes the fragments but oversegments them, discarding the coral body; INSID3+CC does not suffer from this. We hypothesize that INSID3's threshold τ_{fg} is scale-dependent and should be chosen per crop: a lower τ_{fg} suits targets that occupy a large share of the image, and a higher τ_{fg} suits tiny ones. Because we fix a single τ_{fg} across all crops, Foveate oversegments where the target is large and undersegments where it is tiny — which also explains why the single-pass INSID3+CC can beat Foveate on corals.



(a) Ground truth **(b)** Prediction

(c) Cascade

Fig. 4: One cross-image sample from the polyps set. a) shows the ground truth annotations, while b) shows the predictions. c) shows the cascade as a connected graph, where each node represents one crop at a certain depth on the x axis and its crop similarity on the y axis.

The polyps set (Fig. 4) shows the same pattern: a few polyps are segmented correctly, some are oversegmented, but the majority are missed and there are several false positives. The partial coverage again stems from the fixed τ_{fg}, while the missed instances stem from the faulty $k = 2$-means split (Sec. 4.5). On a large crop such as a whole coral, the split produces a thin outer band around an inner disc, both similar in size to the parent: the outer band leaves the crop essentially unchanged, and the inner disc is only a marginal zoom. Both converge below the crop similarity floor τ_C and are rejected. Finally, the false positives arise because we did not tune τ_C: Fig. 4c suggests that a higher crop similarity threshold would filter out many predictions, potentially including the false positives. This is yet another parameter to tune, which we leave for future work.

Inspecting the cascades, we can see that crop similarity uniformly increases with each new crop/split until we oversegment an instance, demonstrating that our hypothesis about zooming in on crops holds. Additionally, we can see a difference between the intra-image (see Fig. 3c) and the cross-image setting (see Fig. 4c). For the former, we have a single leaf reaching near perfect crop similarity. This is the exemplar being re-identified. Non-exemplar crops reach only a crop similarity range of 0.5 to 0.7.

5 Discussion

Our results support the core hypothesis: A recursive zoom lets a small, frozen backbone reach instances that a single pass misses. However, our approach is still in development and exposes a few concrete failure modes. In the following, we discuss our results and implied limitations.

The main bottleneck is Where*, not the stopping signal.* Our stopping-signal validation (Sec. 4.2, Fig. 2) shows that the CLS re-identification score $g(\cdot)$ is

well-behaved on both datasets: it rises as a crop tightens onto an instance, peaks at the true-instance scale, and falls under over-zoom. The recursion criterion is therefore sound in isolation. What caps end-to-end quality is the *Where* step: the emitted masks are bounded by INSID3's foreground quality, and given ground-truth masks the peak sharpens markedly. Foveate's ceiling is thus set by the foreground extractor, not by the zoom-and-stop idea, so a stronger training-free *Where* is a promising next step.

Scale-dependent thresholds. A consistent limitation is the fixed foreground threshold τ_{fg}. Because the concept occupies a very different fraction of the crop at the root than in a leaf, a single τ_{fg} cannot be right at all depths, causing over- and undersegmentation (Sec. 4.7). A promising direction is to schedule τ_{fg} as a function of crop scale or foreground fraction, or to calibrate it per crop from the exemplar bank rather than fixing it globally.

Splitting converged crops. Our $k = 2$-means split does not individuate touching instances; it partitions coral crops into inner/outer regions that both survive re-identification. Replacing it with a component-aware or spectral split on the masked features, or with a split criterion that explicitly rewards balanced, exemplar-consistent children, is an impactful next step, especially for dense concepts such as polyps. We argue that splitting is a necessary step for actually individuating clumps of instances, although our results show that performance degrades.

Untuned stopping. We did not tune the crop similarity floor τ_C as a gate on emitted leaves. Fig. 4c indicates that a higher threshold would suppress many false positives, suggesting easy gains from a principled stopping rule.

Scope and validity. These are early results, and our evaluation scope is narrow in ways that bound what can be concluded from them. Both benchmarks come from a single laboratory setup and, by construction, share the same images: they differ only in the target class, so together they probe *granularity* rather than domain shift, and they say nothing about non-convex, overlapping, or in-situ targets. We further evaluate a single backbone (DINOv3 ViT-S/16) at a single input resolution, compare against two baselines, and report single runs without confidence intervals. Hyperparameters are tuned on the intra-image set, so iAP is optimistic and cAP is the number to read for generalization. Finally, SAM 3 with the image-concat workaround remains the strongest method on corals. Our claims are therefore about feasibility and efficiency (competitive coverage at $\sim 40\times$ fewer parameters and a fraction of the memory) rather than state-of-the-art accuracy; establishing generality requires data from other domains and imaging setups, multiple backbones, and a larger sweep.

Broader impact for ecology. Instance annotation of coral fragments and polyps is a central bottleneck in reef-health studies, forcing ecologists to trade accuracy for speed. A training-free, in-context method targets this directly: a single

exemplar segments every instance of a concept without per-domain retraining, so a new morphology or imaging setup needs only as few as one exemplar. The small, frozen backbone deploys on the modest hardware common in field stations and resource-limited labs. Without a decoder head, Foveate is flexible towards label space changes (e.g. adding a new morphology class), where traditional methods would require complete retraining. Additionally, it is interpretable by construction letting an expert audit why a region was accepted before trusting the measurement. Each mask traces to its most similar exemplar and the full cascade is inspectable (which is how we diagnosed the k-means failure, Sec. 4.7) unlike common black box approaches.

6 Conclusion

We presented **Foveate**, a training-free, in-context, recursive method for instance segmentation built entirely on frozen DINOv3 features. By repeatedly redirecting the encoder's attention to concept-similar regions and re-embedding tighter crops, Foveate recovers small and densely packed instances that a single-pass exemplar baseline misses, using no fine-tuning, decoder, or class labels. On two custom marine benchmarks it matches or exceeds the single-pass baseline and runs at a fraction of a foundation model's memory footprint, though polyps remain hard for every method. Our analysis pinpoints the key obstacles — foreground quality and the split step — and future work will address these alongside more comprehensive benchmarking.

Acknowledgements

This research is part of the Computational Sustainability & Technology project area[5], and has received funding from the Helmholtz Institute for Functional Marine Biodiversity HIFMB, a collaboration between the Alfred Wegener-Institute, Helmholtz-Center for Polar and Marine Research AWI, and the Carl-von-Ossietzky University Oldenburg, initially funded by the Ministry of Science and Culture of Lower Saxony (MWK) (project "An Expert-in-the-Loop ML Tool for Coral Morphometric Analysis") and the german Federal Ministry of Research, Technology and Space (BMFTR) under grant number 16IW26002 (AMENABLE). The data (image repository) was created by HIFMB and is part of the project "Global Search for Genetic Regulators of Coral Resilience to Thermal Stress" supported by grants from the National Oceanographic and Atmospheric Administration (NA20NMF4820320-T1-01) and the Paul G Allen Family Foundation (PGAFF).

[5] `https://cst.dfki.de/`

Bibliography

[1] Archit, A., Freckmann, L., Nair, S., Khalid, N., Hilt, P., Rajashekar, V., Freitag, M., Teuber, C., Spitzner, M., Tapia Contreras, C., Buckley, G., von Haaren, S., Gupta, S., Grade, M., Wirth, M., Schneider, G., Dengel, A., Ahmed, S., Pape, C.: Segment Anything for Microscopy. Nature Methods **22**(3), 579–591 (Mar 2025). https://doi.org/10.1038/s41592-024-02580-4, https://www.nature.com/articles/s41592-024-02580-4

[2] Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (Apr 2023). https://doi.org/10.48550/arXiv.2301.08243, http://arxiv.org/abs/2301.08243, arXiv:2301.08243 [cs]

[3] Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Mojtaba, Komeili, Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Hogan, F.R., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., Szafraniec, M., Krishnakumar, K., Li, Y., Ma, X., Chandar, S., Meier, F., LeCun, Y., Rabbat, M., Ballas, N.: V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning (Jun 2025). https://doi.org/10.48550/arXiv.2506.09985, http://arxiv.org/abs/2506.09985, arXiv:2506.09985 [cs]

[4] Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: SAM 3: Segment Anything with Concepts (Nov 2025). https://doi.org/10.48550/arXiv.2511.16719

[5] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. pp. 9650–9660 (2021), https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper

[6] Couch, C.S., Burns, J.H.R., Liu, G., Steward, K., Gutlay, T.N., Kenyon, J., Eakin, C.M., Kosaki, R.K.: Mass coral bleaching due to unprecedented marine heatwave in Papahānaumokuākea Marine National Monument (Northwestern Hawaiian Islands). PLOS ONE **12**(9), e0185121 (Sep 2017). https://doi.org/10.1371/journal.pone.0185121, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0185121

[7] Cuttano, C., Trivigno, G., Averta, G., Masone, C.: SANSA: Unleashing the Hidden Semantics in SAM2 for Few-Shot Segmentation (Nov 2025). https://doi.org/10.48550/arXiv.2505.21795, http://arxiv.org/abs/2505.21795, arXiv:2505.21795 [cs]

[8] Cuttano, C., Trivigno, G., Reich, C., Cremers, D., Masone, C., Roth, S.: INSID3: Training-Free In-Context Segmentation with DINOv3 (Mar 2026). https://doi.org/10.48550/arXiv.2603.28480, http://arxiv.org/abs/2603.28480, arXiv:2603.28480 [cs.CV]

[9] Humanes, A., Bay, L., Riginos, C., Aguirre, J.D., Angilletta, M.J., Aranda, M., Baker, A.C., Baria-Rodriguez, M.V., Baums, I.B., Bhagooli, R., Duarte, C.M., Guest, J.R., Hendry, A.P., Kenkel, C.D., Lasky, J.R., Mead, D., van Oppen, M.J.H., Peixoto, R.S., Le Port, A., Razak, T.B., Reusch, T.B.H., Scharfenstein, H., Schmidt-Roach, S., Schoepf, V., Suggett, D.J., Voolstra, C.R., Wilson, A.J., Ortiz, J.C.: Accelerating coral assisted evolution to keep pace with climate change. Nature Reviews Biodiversity **2**(5), 306–321 (May 2026). https://doi.org/10.1038/s44358-026-00147-z, https://www.nature.com/articles/s44358-026-00147-z

[10] Jo, S., Lee, S.J., Hong, S., Gang, Y., Kim, H., Seo, H., Kim, K.: One Click per Cell Type Suffices: Training-free Group Interaction for Cell Instance Segmentation (May 2026). https://doi.org/10.48550/arXiv.2605.29429, http://arxiv.org/abs/2605.29429, arXiv:2605.29429 [cs.CV]

[11] Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic Segmentation (Apr 2019). https://doi.org/10.48550/arXiv.1801.00868, http://arxiv.org/abs/1801.00868, arXiv:1801.00868 [cs]

[12] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything (Apr 2023). https://doi.org/10.48550/arXiv.2304.02643, http://arxiv.org/abs/2304.02643, arXiv:2304.02643 [cs]

[13] Li, H., Lian, S., Li, Z., Cong, R., Li, C., Yang, L.T., Zhang, W., Kwong, S.: Advancing Marine Research: UWSAM Framework and UIIS10K Dataset for Precise Underwater Instance Segmentation (Sep 2025). https://doi.org/10.48550/arXiv.2505.15581, http://arxiv.org/abs/2505.15581, arXiv:2505.15581 [cs.CV]

[14] Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft COCO: Common Objects in Context (Feb 2015). https://doi.org/10.48550/arXiv.1405.0312, http://arxiv.org/abs/1405.0312, arXiv:1405.0312 [cs]

[15] Liu, A., Xue, R., Cao, X.R., Shen, Y., Lu, Y., Li, X., Chen, Q., Chen, J.: MedSAM3: Delving into Segment Anything with Medical Concepts (Nov 2025). https://doi.org/10.48550/arXiv.2511.19046, http://arxiv.org/abs/2511.19046, arXiv:2511.19046 [cs.CV]

[16] Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching (Jan 2024). https://doi.org/10.48550/arXiv.2305.13310, http://arxiv.org/abs/2305.13310, arXiv:2305.13310 [cs]

[17] Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment Anything in Medical Images. Nature Communications **15**(1), 654 (Jan 2024). https://doi.org/10.1038/s41467-024-44824-z, http://arxiv.org/abs/2304.12306, arXiv:2304.12306 [eess.IV]

[18] Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: MedSAM2: Segment Anything in 3D Medical Images and Videos (Apr 2025), https://arxiv.org/abs/2504.03600v1

[19] Meng, L., Lan, S., Li, H., Alvarez, J.M., Wu, Z., Jiang, Y.G.: SEGIC: Unleashing the Emergent Correspondence for In-Context Segmentation (Jul 2024). https://doi.org/10.48550/arXiv.2311.14671, http://arxiv.org/abs/2311.14671, arXiv:2311.14671 [cs.CV]

[20] Messmer, V., Jones, G.P., Munday, P.L., Holbrook, S.J., Schmitt, R.J., Brooks, A.J.: Habitat biodiversity as a determinant of fish community structure on coral reefs. Ecology **92**(12), 2285–2298 (2011). https://doi.org/10.1890/11-0037.1, https://onlinelibrary.wiley.com/doi/abs/10.1890/11-0037.1, _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/11-0037.1

[21] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision (Feb 2024). https://doi.org/10.48550/arXiv.2304.07193, http://arxiv.org/abs/2304.07193, arXiv:2304.07193 [cs.CV]

[22] Pachitariu, M., Rariden, M., Stringer, C.: Cellpose-SAM: superhuman generalization for cellular segmentation (May 2025). https://doi.org/10.1101/2025.04.28.651001, https://www.biorxiv.org/content/10.1101/2025.04.28.651001v1, pages: 2025.04.28.651001 Section: New Results

[23] Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment Anything in Images and Videos (Oct 2024). https://doi.org/10.48550/arXiv.2408.00714, http://arxiv.org/abs/2408.00714, arXiv:2408.00714 [cs.CV]

[24] Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (Aug 2025). https://doi.org/10.48550/arXiv.2508.10104, http://arxiv.org/abs/2508.10104, arXiv:2508.10104 [cs.CV]

[25] Uelwer, T., Robine, J., Wagner, S.S., Höftmann, M., Upschulte, E., Konietzny, S., Behrendt, M., Harmeling, S.: A survey on self-supervised methods for visual representation learning. Machine Learning **114**(4), 111 (Mar 2025). https://doi.org/10.1007/s10994-024-06708-7, https://doi.org/10.1007/s10994-024-06708-7

[26] Wilson, S.K., Robinson, J.P.W., Chong-Seng, K., Robinson, J., Graham, N.A.J.: Boom and bust of keystone structure on coral reefs. Coral Reefs

38(4), 625–635 (Aug 2019). https://doi.org/10.1007/s00338-019-01818-4, https://doi.org/10.1007/s00338-019-01818-4

Table 4: Best hyperparameter values for all methods, swept in the multi-exemplar setting. τ_{fg}: INSID3 foreground-granularity threshold; AggT: INSID3 aggregation threshold; t: SAM 3 confidence threshold. Slashes (/) mark parameters that do not apply. INSID3+CC did not find polyps with any configuration, hence we do not report this here.

	Corals			Polyps		
Method	τ_{fg}	AggT	t	τ_{fg}	AggT	t
SAM 3 (4)						
text prompt	/	/	0.3	/	/	0.1
image concat	/	/	0.3	/	/	0.1
INSID3+CC (single-pass)	0.4	0.3	/	/	/	/
Foveate (ours)						
w/ k-means split	0.85	0.0	/	0.975	0.0	/
w/o split	0.8	0.0	/	0.975	0.0	/

A Hyperparameters

Table 4 lists the best hyperparameter configuration found for each method in the multi-exemplar setting.