This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

INSID3: Training-Free In-Context Segmentation with DINOv3

Claudia Cuttano[*1,2] Gabriele Trivigno[*1] Christoph Reich[2,3,5,6]
Daniel Cremers[3,5,6] Carlo Masone[1] Stefan Roth[2,4,5]
[1]Politecnico di Torino [2]TU Darmstadt [3]TU Munich [4]hessian.AI [5]ELIZA [6]MCML [*]equal contribution
https://visinf.github.io/INSID3

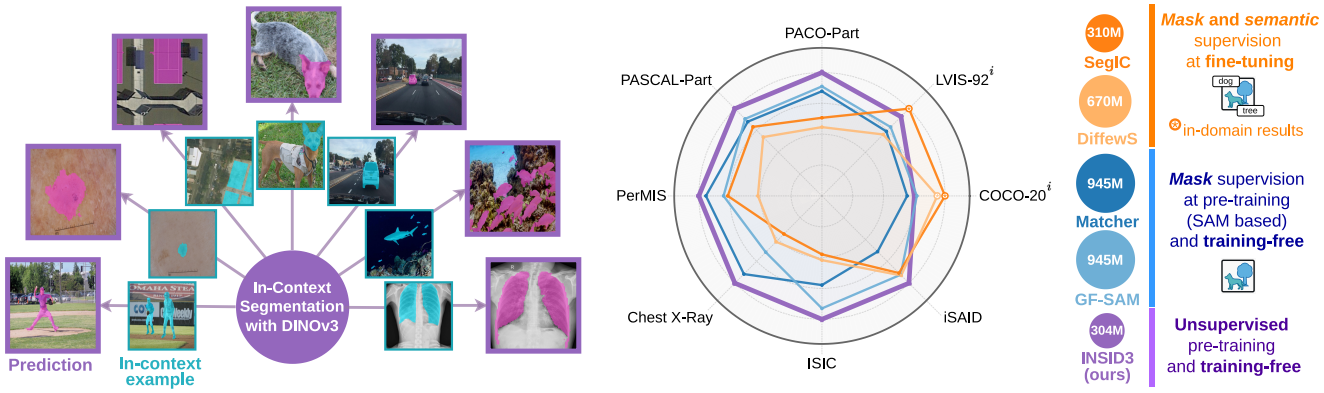


Figure 1. **Results and overview of INSID3, our training-free in-context segmentation approach**. INSID3 performs in-context segmentation directly from DINOv3 [56] features, without any decoder, fine-tuning, or model composition. *(left)* A single annotated example guides the model to segment any concept, from object parts to medical images and aerial views. *(right)* Comparing generalization across datasets and segmentation granularities: fine-tuned methods *(orange)* excel in-domain (⊛) but degrade out of distribution, while SAM-based pipelines *(blue)* generalize better but rely on large, multi-stage architectures. INSID3 *(purple)* achieves the strongest generalization with a single backbone, revealing that robust segmentation can emerge directly from the dense self-supervised representations of DINOv3.

Abstract

In-context segmentation (ICS) aims to segment arbitrary concepts, e.g.*, objects, parts, or personalized instances, given one annotated visual examples. Existing work relies on* (i) *fine-tuning vision foundation models (VFMs), which improves in-domain results but harms generalization, or* (ii) *combines multiple frozen VFMs, which preserves generalization but yields architectural complexity and fixed segmentation granularities. We revisit ICS from a minimalist perspective and ask: Can a single self-supervised backbone support both semantic matching and segmentation, without any supervision or auxiliary models? We show that scaled-up dense self-supervised features from DINOv3 exhibit strong spatial structure and semantic correspondence. We introduce INSID3, a training-free approach that* segments concepts *at varying granularities only from frozen DINOv3 features, given an in-context example. INSID3 achieves state-of-the-art results across one-shot semantic, part, and personalized segmentation, outperforming previous work by +7.5 % mIoU, while using 3× fewer parameters and without any mask or category-level supervision.*

1. Introduction

Understanding visual scenes is a fundamental task with applications in autonomous driving [11, 31], robotics [18], augmented reality [34], or medical image analysis [62]. In-context segmentation (ICS) [39, 41, 72] approaches the task of segmenting arbitrary concepts, such as *objects*, *parts*, or *personalized instances* in images, given one or more annotated examples at inference time, *cf*. Fig. 1 *(left)*. This holistic and open-world scene understanding task requires adaptability to different reference annotations and domains, sharing the spirit of adapting large language models (LLMs) through contextual instructions to novel tasks [3, 9, 48, 58].

ICS requires reliable visual correspondences between annotated reference examples and target images. Previous work showed that such visual correspondences emerge in features of vision foundation models (VFMs) [57, 70]. Based on this, recent work has explored how to endow VFMs with explicit segmentation capabilities. For instance, [38, 41, 74] augment a frozen DINOv2 [47] by training a segmentation decoder on top or fine-tune a diffusion model [52] through episodic training. These approaches aim to

translate implicit visual understanding of VFMs into dense, pixel-level predictions. Although this boosts in-domain results, it requires additional supervision and narrows the model scope to the training distribution (*cf*. Fig. 1, *orange*).

In contrast, recent training-free approaches [39, 69] forego task-specific training, exploiting the complementary strengths of multiple pre-trained components: DINOv2 [47] for robust visual correspondence and SAM [33] for producing accurate masks. By relying purely on pre-trained models, these methods avoid the pitfalls of fine-tuning, achieving stronger generalization (Fig. 1, *blue*). Nevertheless, they need to coordinate multiple VFMs, add significant computational overhead, and cannot fully exploit the intrinsic synergy between correspondence and segmentation.

Overall, existing ICS methods rely explicitly or implicitly on segmentation priors learned through supervision, whether from SAM pre-training or downstream fine-tuning. The recent DINOv3 model [56] may hold the key to changing this. This purely self-supervised VFM, trained on massive-scale image corpora, is explicitly designed to produce dense localized features, unlike its predecessors [7, 47]. Its objective preserves spatial structure, enabling robust region-level grouping (Fig. 2). This urges us to ask if ICS can emerge directly from the DINOv3 representation, without any decoder, fine-tuning, or model composition.

To this end, we propose INSID3 (**In**-context **S**egmentation w**I**th **DINOv3**), a minimalist and training-free approach, relying solely on DINOv3 features. INSID3 operates in three conceptual stages: *(i) Fine-grained clustering* of target image features allows to obtain part-level region candidates (Fig. 2). *(ii) Seed-cluster selection* identifies the most discriminative cluster through cross-image similarity between a prototype of the annotated example(s) and each cluster in the target. Relying on region-level similarity suppresses spurious pixel matches and resolves competition among many candidates. *(iii) Aggregation guided by self-similarity* of DINOv3 features within the target image then merges the seed cluster with other highly affine clusters, producing a spatially coherent mask that recovers the full extent of the prompted concept.

Finally, we uncover a subtle, yet significant limitation of correspondences from DINOv3: feature similarities across unrelated images exhibit systematic activations aligned with absolute spatial coordinates (*e.g*., features from the left side of two images tend to spuriously match regardless of semantics, as shown in Fig. 4). This *positional bias*, likely an effect of the superposition of positional encodings and semantic signals, hinders reliable correspondence reasoning in matching tasks. We propose a simple correction: we estimate the subspace affected by positional bias from a noise image and perform matching only in its orthogonal complement. This lightweight operation improves cross-image matching and, as we show, even generalizes beyond ICS.

Figure 2. **Region-level grouping from DINOv3.** Each pair shows an input image *(left)* and the corresponding clustering map *(right)* obtained by applying agglomerative clustering to dense DINOv3 features. The resulting clusters delineate coherent object- and part-level regions, providing a structured decomposition of the scene.

In summary, we propose INSID3, a principled, minimalist, yet accurate method for in-context segmentation from DINOv3 alone. It is applicable across diverse semantic granularities, *e.g*., from *objects* to *parts*, and demonstrates that emergent segmentation behavior can arise naturally from self-supervision without any training or fine-tuning.

Summarizing, we make the following contributions:

- We are the first to show that *a self-supervised VFM suffices for training-free in-context segmentation*, building on DINOv3's core strengths of robust correspondence and its dense, localized feature structure.
- Despite its simplicity, INSID3 *generalizes better across the board*, from traditional, challenging benchmarks to out-of-domain datasets and part segmentation (Fig. 1, *purple*), outperforming fine-tuned *and* training-free approaches relying on SAM by an average of +7.5 % mIoU.
- We unveil a *positional bias in DINOv3*, which impairs its effectiveness in matching features across images, and present a simple training-free correction that generalizes beyond ICS, achieving gains of up to +6.6 % PCK on the related task of semantic correspondence.

2. Related Work

In-context segmentation (ICS) draws inspiration from LLMs [3, 9, 48, 58], which can be adapted to new tasks given contextual examples. SegGPT [64] and Painter [63] translate this idea to computer vision by training a generalist model to handle multiple segmentation scenarios. Recently, this idea has been revisited in light of the advent of large-scale pre-trained VFMs: Matcher [39] uses an annotated example to perform one-shot *semantic* and *part* segmentation, while PerSAM [72] focuses on one-shot *personalized* segmentation. Although related in spirit to few-shot segmentation [12, 27, 37, 61], which learns from base classes and evaluates on disjoint novel ones defined within each dataset, ICS differs in scope and evaluation. In particular, we refer

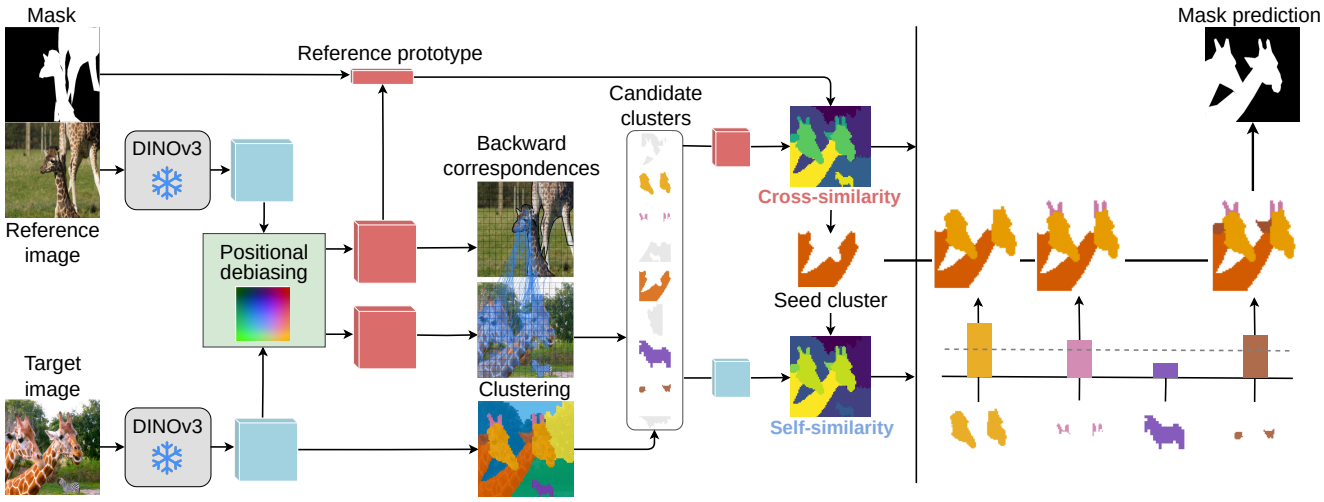


Figure 3. **Overview of INSID3.** We leverage the semantic and spatial structure of DINOv3 to perform in-context segmentation without training or model composition. Dense features from the reference and target images are first debiased to suppress positional bias, improving cross-image matching. The target is then decomposed into coherent regions through agglomerative clustering, providing a structured representation. We retain candidate clusters that match the reference through backward correspondence in the debiased space; a reference prototype derived from the annotated region anchors the *seed cluster* via cross-image similarity. Finally, we combine cross-image similarity, capturing semantic alignment, with self-similarity, measuring the affinity of each cluster to the seed, to form the final mask from the seed.

to ICS as a unified formulation of one-shot *semantic*, *part*, and *personalized* segmentation across different levels of semantic granularity within a single, general-purpose model.

Recent work follows two trends: *Training-free pipelines* [16, 39, 69] combine the semantic understanding of DINOv2 with segmentation priors from SAM, benefiting from strong generalization but inheriting SAM's mask granularity and the computational burden of multi-stage designs. *Supervised methods* [41, 74] aim to unify both capabilities within a single VFM by injecting segmentation functionality via task-specific supervision. SegIC [41] trains a segmentation decoder on top of DINOv2, while DiffewS [74] fine-tunes Stable Diffusion [52]. Such training/fine-tuning couples the model to the training distribution, limiting its flexibility on unseen domains and granularities. In contrast, we address ICS with a single VFM *and* without training.

Dense self-supervised representation learning (SSL) aims to learn dense feature extractors from unlabeled data, enabling a broad range of vision tasks [15, 20]. Initial self-supervised approaches employ image-level pre-text tasks [5, 6, 8, 14, 19, 23, 35, 45], transferring suboptimally to pixel-level prediction [66, 67]. Later work aims to learn localized and discriminative dense features. Emergent properties in ViTs [7] can be uncovered through spatially local objectives [47, 73], spatio-temporal consistency [29], or spatial alignment across views [1, 46]. Localized supervision is also possible through contrastive objectives on region proposals [25, 26], or by predicting the cluster identity of masked tokens [13]. Moreover, SSL features can be refined *a-posteriori* [22, 65] or through limited fine-tuning [32, 53]. Recent efforts distill DINOv2 [47] together with weakly supervised VFMs, *e.g.*, SAM [33] or CLIP [49], to enhance spatial fidelity [2, 24, 51]. Most recently, DINOv3 [56] uses significant data and model scaling, a Gram anchoring objective, and high-resolution post-training to obtain an expressive, dense feature extractor. We show that dense DINOv3 features can be directly leveraged for in-context segmentation without fine-tuning or model composition.

3. In-context Segmentation with INSID3

Our goal is to segment arbitrary concepts, *i.e.*, objects, parts, or personalized instances, given an in-context example, using a frozen DINOv3 encoder without training or model composition. A key property of DINOv3 is the strong self-similarity of its dense features, naturally grouping coherent parts or objects (Fig. 2). However, in-context segmentation also requires establishing correspondences *across* images, which we find affected by a systematic *positional bias*: features from similar positions spuriously match across unrelated images. To address this, we propose a simple, training-free strategy to remove positional components from the features (Sec. 3.1). We use these *debiased features* for cross-image matching, while retaining the original features for intra-image similarity and clustering.

Our approach, named INSID3 and illustrated in Fig. 3, first partitions the target image into semantically coherent regions using self-similarity (Sec. 3.2). Then it identifies the cluster that is most semantically aligned with the reference region through cross-image similarity in the debiased space

(Sec. 3.3). Finally, it expands this seed region by aggregating clusters according to intra-image self-similarity, yielding a complete and coherent segmentation mask (Sec. 3.4).

Task definition. We let $\mathbf{I}^r \in \mathbb{R}^{H\times W\times 3}$ denote the reference image with its binary mask $\mathbf{M}^r \in \{0,1\}^{H\times W}$, and $\mathbf{I}^t \in \mathbb{R}^{H\times W\times 3}$ a target image. We extract dense features from a frozen DINOv3 encoder $\Phi(\cdot)$ [56]:

$$\mathbf{F}^r = \Phi(\mathbf{I}^r), \qquad \mathbf{F}^t = \Phi(\mathbf{I}^t), \tag{1}$$

where $\mathbf{F}^r, \mathbf{F}^t \in \mathbb{R}^{P\times D}$ denote the D-dimensional patch embeddings at resolution $P = H' \times W'$. We let $\Omega = \{1, \dots, P\}$ denote the set of patch indices in $\mathbf{F}^r$ and $\mathbf{F}^t$.

3.1. Unlocking the DINOv3 feature space

Solving the ICS task fundamentally relies on computing robust and reliable feature correspondences between the reference and target images [39]. As a diagnostic tool to evaluate DINOv3's ability to establish reliable correspondences, we compute cross-image similarity to visualize how target patches align with the reference concept. Specifically, given the reference mask $\mathbf{M}^r$ and the set of foreground patch indices[1] $\mathcal{R} = \{j \in \Omega \mid \mathbf{M}^r_j = 1\}$, we compute a reference prototype $\mathbf{p}^r$ and its similarity to each target patch $i \in \Omega$:

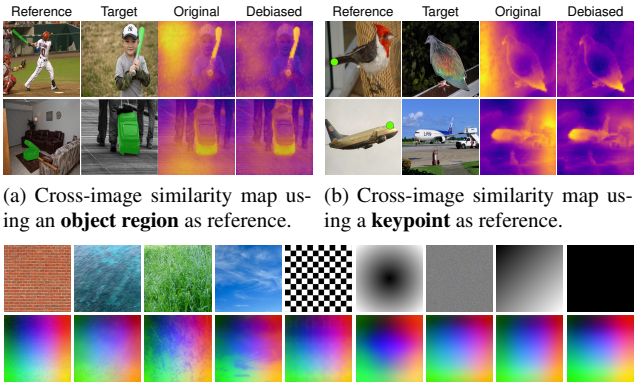
$$\mathbf{p}^r = \frac{1}{|\mathcal{R}|}\sum\nolimits_{j\in\mathcal{R}} \mathbf{F}^r_j, \qquad \text{sim}(i) = \langle \mathbf{F}^t_i, \mathbf{p}^r \rangle. \tag{2}$$

This produces dense similarity maps, indicating how well each target patch aligns with the reference concept. We visualize these maps at two granularity levels: *(i)* at mask level (Fig. 4a), where the reference corresponds to an object, and *(ii)* at keypoint level (Fig. 4b), where the reference is a single annotated keypoint. The resulting similarity maps show that DINOv3 captures meaningful semantic correspondences between reference and target. However, they also exhibit a stable *positional bias*: features at a given position in the reference tend to produce spurious activations at the same position in the target, irrespective of semantics. These false activations typically occur where the target area lacks semantic content (*e.g.*, uniform background regions), suggesting that positional information dominates weak semantic cues. To ground this intuition, Fig. 4c visualizes features from inputs with minimal semantic content: a principal component analysis (PCA) suggests a stable low-dimensional subspace associated with positional signals.

We use this signal as a simple and effective approximation of positional bias, which can be estimated once and removed consistently at inference time. Specifically, we estimate the positional subspace by passing a noise image $\mathbf{I}^{\text{noise}} \sim \mathcal{N}(\mathbf{0}, \mathbf{1}) \in \mathbb{R}^{H\times W\times 3}$ through the encoder:

$$\mathbf{F}^{\text{noise}} = \Phi(\mathbf{I}^{\text{noise}}) \in \mathbb{R}^{P\times D}. \tag{3}$$

[1] In slight abuse of notation, we let $\mathbf{M}^r_j$ denote the j^{th} patch of $\mathbf{M}^r$.



(a) Cross-image similarity map using an **object region** as reference. (b) Cross-image similarity map using a **keypoint** as reference.

(c) **Positional subspace.** PCA on uniform and low-complexity images.

Figure 4. **Positional bias in DINOv3 features.** For both region *(a)* and keypoint *(b)* prompts, similarity maps computed with the original DINOv3 features show structured activations aligned with the reference coordinates, independent of semantics. Our debiased features mitigate this behavior. *(c)* PCA of features from images with low semantic complexity (*e.g.*, noise, flat textures) reveals a stable low-dimensional positional subspace underlying this bias.

We apply singular value decomposition $\mathbf{F}^{\text{noise}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and select the top s right singular vectors $\mathbf{B} = \mathbf{V}_{[:,1:s]}$ as a basis for the positional subspace. We then project both reference and target features onto its orthogonal complement:

$$\tilde{\mathbf{F}}^r = \mathbf{F}^r\big(\mathbf{1}_D - \mathbf{B}\mathbf{B}^\top\big) \qquad \tilde{\mathbf{F}}^t = \mathbf{F}^t\big(\mathbf{1}_D - \mathbf{B}\mathbf{B}^\top\big). \tag{4}$$

The effect of this projection is to suppress positional components: using these debiased features to recompute the similarity map as in Eq. (2) yields activations that are less affected by structured positional bias (*cf.* Fig. 4). In the rest of the paper, we refer to $\tilde{\mathbf{F}}^r, \tilde{\mathbf{F}}^t$ as *debiased features*.

Interestingly, this spatial dependency is markedly weaker in DINOv2; positional correlations are less pronounced and not as easily observable in similarity maps (*cf*. Supp. Material). We hypothesize this positional bias to be a by-product of the stronger local-consistency constraints in DINOv3. Namely, the *Gram anchoring* constrains the covariance matrix of patch embeddings, encouraging global statistics of features to remain stable throughout training. While improving spatial consistency, this objective may inadvertently amplify absolute spatial correlations, resulting in residual positional bias when semantic content is weak.

Unlocked feature space. We exploit the complementary nature of DINOv3 features by using: *(i)* our *debiased features* for cross-image semantic matching, where positional signals are harmful, and *(ii) original features* for intra-image grouping, where spatial structure is helpful.

3.2. Fine-grained clustering

The first step of our ICS pipeline is to partition the target image into semantically coherent regions. The dense feature maps of DINOv3 exhibit strong local consistency: As

shown by [56], patches belonging to the same object or part tend to have highly similar embeddings. We leverage this to group image regions in an unsupervised manner.

While K-means clustering has been widely used in self-supervised representation learning [6, 22, 40, 47], it requires predefining the number of clusters, which is ill-suited for the open-world nature and variable granularity of ICS. Density-based methods, *e.g.*, DBSCAN [17], struggle in high-dimensional feature spaces where the notion of density becomes unreliable [50, 55], and typically require dimensionality reduction. Instead, we adopt agglomerative clustering [43], which progressively merges locally similar features in a bottom-up manner, naturally aligning with the spatial smoothness of DINOv3. A single threshold hyperparameter τ provides intuitive control over the resulting granularity without fixing a predefined number of regions.

Concretely, we partition the (original) target patch embeddings $\mathbf{F}^t$ into K disjoint spatial regions via iterative agglomerative clustering, yielding clusters $\{\mathcal{G}_1, \dots, \mathcal{G}_K\}$ such that

$$\bigcup_{k=1}^{K} \mathcal{G}_k = \Omega, \qquad \mathcal{G}_i \cap \mathcal{G}_j = \emptyset \quad \forall i \neq j. \tag{5}$$

As shown in Fig. 2, this unsupervised approach produces spatially coherent clusters that provide a strong structural representation of the image.

3.3. Seed-cluster selection

Having partitioned the target image into semantically coherent regions, we identify the cluster that best corresponds to the reference region. We do this in two stages: *candidate localization* and *seed-cluster selection*.

Candidate localization. Directly correlating the reference prototype with all target patches, as in Fig. 4a, often produces broad activations on related concepts, even when the reference depicts only a specific part: *e.g.*, the prototype of a *person head* may trigger responses over the full person.

To adapt matching at the correct level of granularity, we instead compute *backward* correspondences, *i.e.* for each target patch i, we find its most similar reference patch j:

$$\mathrm{NN}(i) = \arg\max_{j \in \Omega} \langle \tilde{\mathbf{F}}^t_i, \tilde{\mathbf{F}}^r_j \rangle, \tag{6}$$

Backward matching of target patches allows us to *implicitly leverage unannotated negatives* in the reference image. By retaining only target patches whose nearest neighbor in the reference falls within the support mask, we obtain a filtering mechanism that conservatively estimates the set $\mathcal{C}_{\mathrm{NN}}$ of target patches in which the reference concept may appear as

$$\mathcal{C}_{\mathrm{NN}} = \big\{ i \,\big|\, \mathbf{M}^r_{\mathrm{NN}(i)} = 1 \big\}. \tag{7}$$

Restricting the precomputed clusters $\{\mathcal{G}_k\}$ to those that overlap with $\mathcal{C}_{\mathrm{NN}}$ yields the subset of candidate clusters

$$\mathcal{C}_{\mathrm{cand}} = \big\{ \mathcal{G}_k \,\big|\, \mathcal{G}_k \cap \mathcal{C}_{\mathrm{NN}} \neq \emptyset \big\}. \tag{8}$$

Seed selection. We compute prototypes in the debiased feature space for both the candidate clusters $\mathcal{G}_k \in \mathcal{C}_{\mathrm{cand}}$ and the annotated reference region:

$$\tilde{\mathbf{p}}^t_k = \frac{1}{|\mathcal{G}_k|} \sum_{i \in \mathcal{G}_k} \tilde{\mathbf{F}}^t_i \qquad \tilde{\mathbf{p}}^r = \frac{1}{|\mathcal{R}|} \sum_{j \in \mathcal{R}} \tilde{\mathbf{F}}^r_j. \tag{9}$$

We then compute a cross-image similarity score

$$s^{\mathrm{cross}}_k = \langle \tilde{\mathbf{p}}^t_k, \tilde{\mathbf{p}}^r \rangle, \tag{10}$$

measuring how well each candidate $\mathcal{G}_k$ aligns semantically with the reference. The final seed cluster is selected as

$$\mathcal{G}^* = \arg\max_{\mathcal{G}_k \in \mathcal{C}_{\mathrm{cand}}} s^{\mathrm{cross}}_k, \tag{11}$$

corresponding to the target region that is most semantically aligned with the reference at the correct part granularity.

3.4. Cluster aggregation

The seed cluster $\mathcal{G}^*$ provides a strong but typically *partial* localization of the semantic concept in the target, often covering only the most discriminative part of the concept, such as a person's head or the neck of a giraffe (*cf*. Fig. 3). To recover the full extent of the concept, we evaluate all remaining candidate clusters to decide which should be merged. Intuitively, the cross-image similarity score s^{cross}_k (Eq. 10), reflects how semantically close candidate clusters $\mathcal{G}_k$ are to the reference. However, relying solely on cross-image similarity can be unreliable under occlusions or viewpoint changes, where semantically relevant regions may appear dissimilar. Therefore, we propose to complement *semantic alignment* (across images) with *structural coherence* (within the target image). Specifically, we exploit a key property of DINOv3 [56]: *its features exhibit strong self-similarity within the same image*. Hence, clusters belonging to the same concept tend to lie close in feature space. For each candidate cluster, we thus compute its similarity to the seed in the *original* feature space as

$$\bar{\mathbf{p}}^t_k = \frac{1}{|\mathcal{G}_k|} \sum_{i \in \mathcal{G}_k} \mathbf{F}^t_i, \qquad s^{\mathrm{intra}}_k = \langle \bar{\mathbf{p}}^t_k, \bar{\mathbf{p}}^t_* \rangle, \tag{12}$$

where $\bar{\mathbf{p}}^t_*$ denotes the prototype of the seed cluster $\mathcal{G}^*$.

Final aggregation. We combine semantic alignment and structural coherence through a multiplicative score, which favors clusters that are simultaneously semantically aligned with the reference and structurally consistent with the seed region. The final mask is obtained by merging the seed cluster with all candidate clusters whose combined score exceeds a similarity threshold α:

$$S_k = s^{\mathrm{cross}}_k \cdot s^{\mathrm{intra}}_k \tag{13}$$

$$\mathcal{M}_{\mathrm{final}} = \mathcal{G}^* \cup \big\{ \mathcal{G}_k \in \mathcal{C}_{\mathrm{cand}} \,\big|\, S_k \geq \alpha \big\}. \tag{14}$$

Table 1. **Comparison of INSID3 (mIoU in %, ↑) on one-shot semantic, part, and personalized segmentation.** State-of-the-art methods are grouped into task-specific fine-tuning and training-free approaches. Previous training-free methods rely on SAM, pre-trained with mask-level supervision, whereas INSID3 uses only frozen self-supervised DINOv3 features. Gray indicates the model was trained on the corresponding train split of the dataset; best results **bold**, 2nd best underlined. † denotes a GF-SAM variant using DINOv3 features.

			Semantic						Part		Personalized	
Method	**Encoder**	**#Param**	**LVIS-92^i**	**COCO-20^i**	**ISIC**	**SUIM**	**iSAID**	**X-Ray**	**PASCAL**	**PACO**	**PerMIS**	**Avg**
Task-specific fine-tuning: *Semantic + mask supervision*												
Painter [63]	ViT	354 M	10.5	33.1	–	–	–	–	30.4	14.1	–	–
SegGPT [64]	ViT	354 M	18.6	56.1	37.5	34.9	30.9	**87.5**	35.8	13.5	18.7	37.1
SINE [38]	DINOv2	373 M	31.2	64.5	25.8	50.7	38.3	39.8	36.2	23.3	42.5	39.1
DiffewS [74]	Stable Diffusion	890 M	31.4	71.3	27.8	48.9	47.5	41.6	34.0	22.8	35.2	40.1
SegIC [41]	DINOv2	310 M	44.6	76.1	25.3	52.5	46.1	34.5	39.9	25.9	51.8	44.1
SegIC (COCO) [41]	DINOv2	310 M	35.7	75.6	22.5	52.9	40.8	30.8	38.6	25.1	44.9	40.8
Training free: *Mask-supervised pre-training*												
PerSAM [72]	SAM	640 M	11.5	23.0	23.9	28.7	19.2	31.7	32.5	22.5	48.6	26.8
Matcher [39]	DINOv2 + SAM	945 M	33.0	52.7	38.6	44.1	33.3	70.8	42.9	34.7	63.8	46.0
GF-SAM [69]	DINOv2 + SAM	945 M	35.2	**58.7**	48.7	53.1	47.1	51.0	44.5	36.3	54.1	47.6
GF-SAM† [69]	DINOv3 + SAM	945 M	31.8	54.8	50.9	50.5	46.7	56.1	44.9	34.4	52.6	47.0
↳ + *our debias*	DINOv3 + SAM	945 M	34.6	55.9	51.8	52.9	47.6	60.0	46.2	36.1	54.5	48.8
Training free: *Unsupervised pre-training*												
INSID3 *(ours)*	DINOv3	**304 M**	**41.8**	57.6	**54.4**	**54.9**	**52.1**	78.8	**50.5**	**38.7**	**67.0**	**55.1**

4. Experiments

We evaluate INSID3 on one-shot *semantic*, *part*, and *personalized* segmentation. In each setting, a single annotated reference mask is provided, and the model is tasked with segmenting the corresponding concept in the target image: *(1) semantic* – all instances of a given class (*e.g.*, "dog"); *(2) part* – same object part (*e.g.*, "dog ear"); *(3) personalized* – same object instance (*e.g.*, "my dog").

For **one-shot *semantic* segmentation**, we use six datasets across a range of imaging scenarios: COCO-20^i [44] with 80 object categories; LVIS-92^i [39] with 920 categories and a strong long-tail distribution; ISIC2018 [10, 59] for skin lesion segmentation; Chest X-Ray [4, 30], an X-ray dataset of lung screening; iSAID-5^i [68], a remote sensing dataset with 15 categories; and SUIM [28] with underwater imagery and 8 categories. For **one-shot *part* segmentation**, we use PASCAL-Part [39], providing 56 object parts across 15 categories, and PACO-Part [39] with 303 object parts from 75 categories. For **one-shot *personalized* segmentation**, we use PerMIS [54], covering 16 categories.

Implementation details. We adopt the Large version of the DINOv3 [56] encoder. Input images are resized to 1024 × 1024, following SAM-based approaches [39, 69, 72]. The final segmentation masks are predicted at patch resolution: we bilinearly interpolate them to original resolution, and apply mask refinement with a CRF [36], following [21, 22, 40, 60]. We employ agglomerative clustering [43] with $\tau = 0.6$ and set the cluster aggregation threshold to $\alpha = 0.2$. See Supplementary Material for more details.

4.1. Main results

Baselines. Table 1 compares against state-of-the-art ICS baselines in terms of mean Intersection-over-Union (mIoU). The primary baselines are *training-free methods*, specifically PerSAM [72], Matcher [39], and GF-SAM [69]. For GF-SAM, the strongest training-free baseline, we also include a variant in which DINOv2 is replaced with DINOv3 and a version with our feature debiasing to ensure a fair comparison. We also report *task-specific fine-tuning* methods such as SegIC [41] and DiffewS [74], which leverage semantic and mask supervision. For SegIC, we also report a version trained only on COCO. While these operate under a different supervision regime, they provide an upper reference point for in-domain accuracy. We emphasize that **INSID3** is the only method in Tab. 1 that uses no supervision (neither during pre-training nor fine-tuning) and operates solely on the self-supervised DINOv3 backbone.

One-shot semantic segmentation. Table 1 shows that INSID3 consistently outperforms training-free SAM-based pipelines, with gains over GF-SAM of +6.6 % pts. mIoU on LVIS-92^i, +5.7 % pts. on ISIC, +1.8 % pts. on SUIM, and +27.8 % pts. on Chest X-ray, *etc*. Upgrading GF-SAM from DINOv2 to DINOv3 yields comparable performance, as its method relies only on sparse matched points to prompt SAM, thus discarding most of the information in DINOv3 dense features. However, our debiasing helps GF-SAM with DINOv3 (pts. mIoU on average). In contrast, INSID3 performs estimation and segmentation in a unified space, achieving both higher accuracy and lower architec-

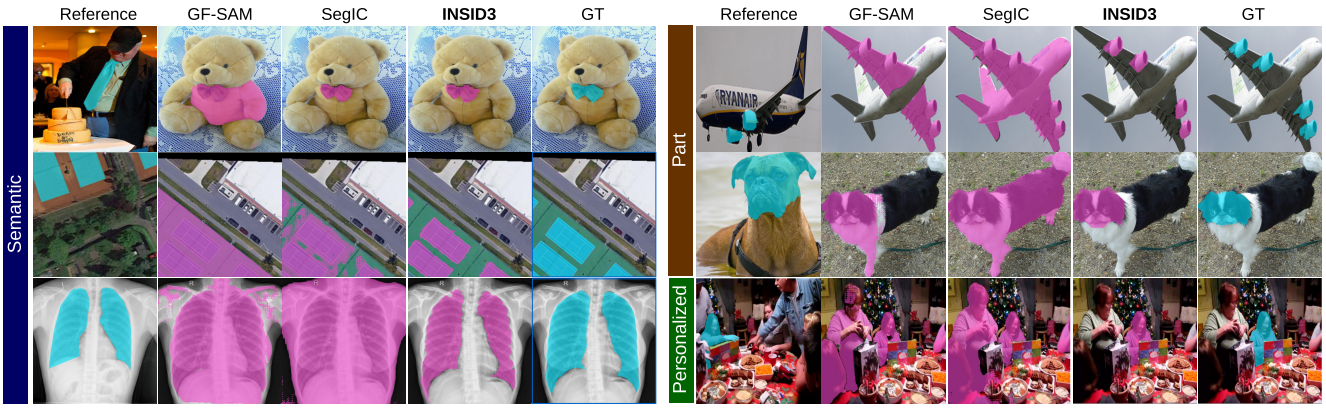


Figure 5. Comparison of **INSID3** with GF-SAM [69] and SegIC [41] on one-shot semantic *(left)*, part *(top right)*, and personalized *(bottom right)* segmentation. **SegIC** performs well in-domain but struggles to generalize across domains and part granularity, reflecting its limited flexibility beyond the training distribution. **GF-SAM**, relying on the strong segmentation priors of SAM [33], produces high-quality masks; however, the decoupled mechanism between correspondence and segmentation often leads to over- or under-segmentation. **INSID3**, despite relying solely on self-supervised features, achieves precise localization and competitive mask quality.

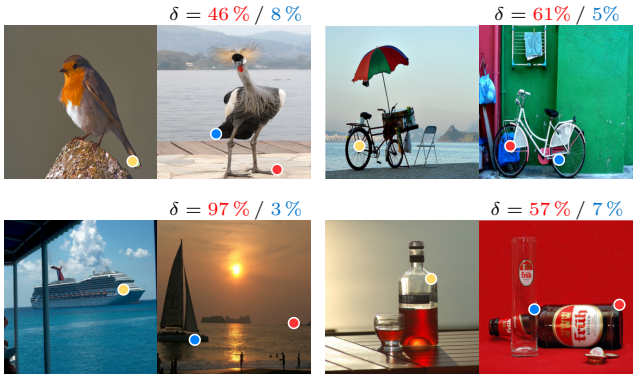
tural complexity (304 M *vs*. 945 M parameters). Interestingly, fine-tuned methods achieve strong in-domain results (*e.g*., SegIC reaches 76.1 % mIoU on COCO-20^i) but drop sharply on other datasets, reflecting the inherent trade-off of specialization (*e.g*. −6 % w.r.t. INSID3 on iSAID). Qualitative results in Fig. 5 *(left)* show that INSID3 produces surprisingly clean segmentation masks directly from DINOv3 features, without any decoder or task-specific supervision.

One-shot part segmentation. INSID3 achieves significant improvements over existing baselines on also on part segmentation (*cf*. Tab. 1). It outperforms GF-SAM by +6.0 % pts. mIoU on PASCAL-Part and +2.4 % pts. mIoU on PACO-Part. Two-stage pipelines frequently over- or under-cover object parts due to fixed mask priors, inherited through fine-tuning or from SAM. In contrast, INSID3 better exploits the reference signal throughout the pipeline, enabling flexible and accurate part-level predictions. Compared to fine-tuned approaches, INSID3 outperforms SegIC and DiffewS by +10.6 % / +16.5 % pts. on PASCAL-Part and +12.8 % / +15.9 % pts. on PACO-Part. As illustrated in Fig. 5 *(top right)*, INSID3 produces part masks that better preserve object structure and segmentation granularity.

One-shot personalized segmentation. INSID3 achieves the best results on PerMIS, reaching 67.0 % mIoU and surpassing GF-SAM by +12.9 % pts., SegIC by +15.2 % pts., and DiffewS by +31.8 % pts. This task is particularly challenging due to the presence of multiple visually similar distractor instances. Unlike previous work, which relies solely on positive activations from the reference and tends to segment all semantically related objects, INSID3 additionally exploits negative evidence through backward correspondences (Sec. 3.3) to suppress irrelevant regions. Figure 5 *(bottom right)* shows that this leads to accurate instance selection, even in the presence of visual ambiguity.

Table 2. **Semantic correspondence on SPair-71k** (PCK@T in %, ↑). Comparison across DINOv3 backbones, w/ and w/o debiasing.

	Small		Base		Large	
T	*original*	*debias*	*original*	*debias*	*original*	*debias*
0.05	26.8	**27.9**	29.2	**32.3**	32.7	**33.6**
0.10	43.8	**45.7**	45.0	**50.0**	50.6	**52.0**
0.15	53.2	**55.6**	54.0	**59.9**	60.3	**62.1**
0.20	59.8	**62.6**	59.8	**66.4**	66.4	**68.6**



● source keypoint; ● prediction (original DINOv3); ● prediction (debiased). δ is the relative error w.r.t. GT.

Figure 6. **Qualitative examples on SPair-71k** with DINOv3-L.

4.2. On the positional bias of DINOv3 features

While we focus on ICS, the positional bias we uncover in DINOv3 has broader implications for tasks that rely on semantic image alignment without spatial priors. A representative example is *semantic correspondence* [57, 70, 71], which evaluates how well dense features can localize semantically corresponding points across different images. We evaluate our debiasing strategy on SPair-71k [42], the

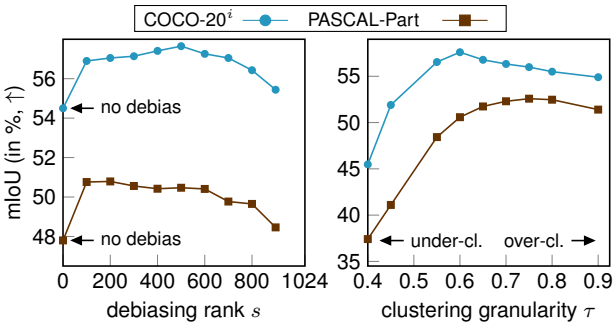


Figure 7. **INSID3 hyperparameters.** We study the effect of the debiasing rank s *(left)* and the clustering granularity τ *(right)* on semantic (COCO-20^i) and part (PASCAL-Part) segmentation.

standard benchmark for this task. Following common evaluation protocol [57, 70], for each source keypoint, we compute cosine similarity to all target patch tokens and select the most similar one. Accuracy is measured using PCK@T (% of Correct Keypoints within a normalized distance T).

Quantitatively, Tab. 2 shows that debiasing leads to consistent gains of +0.9–6.6 PCK across all model sizes. These results show that our positional debiasing acts as a simple training-free correction that improves the reliability of DINOv3 features for tasks requiring semantic alignment across images. Qualitatively, Fig. 6 compares predictions obtained with the original DINOv3 features *(red)* and our debiased features *(blue)*. Predictions based on original features are influenced by a mix of semantic and positional cues, resulting in systematic errors. In contrast, debiased features align closely with the correct semantic location, yielding more semantically grounded matches.

4.3. Ablation study

We conduct ablations on COCO-20^i and PASCAL-Part. For more analyses, please refer to the Supplementary Material.

Debiased feature space. We study the influence of the *rank* s of the estimated positional subspace $\mathbf{B}$ on our proposed *debiasing* of DINOv3 from Eq. (4). Fig. 7 *(left)* shows a stable trend, improving over the *no debiasing* baseline (*i.e.*, $s = 0$), indicating that the removed positional subspace does not carry semantically meaningful information. Beyond an intermediate range, the gain saturates and eventually reverses once too many subspace dimensions are being removed. We fix s to 500 across all datasets, which yields consistent gains (+3.1 % on COCO, +2.7 % on PASCAL and up to +6.6 % PCK on SPair-71k, *cf*. Tab. 2).

Cluster granularity. Next, we analyze the impact of the *similarity threshold* τ of agglomerative clustering (Sec. 3.2). Fig. 7 *(right)* shows that finer partitions resulting from larger τ are beneficial for part-level tasks. Conversely, over-clustering is detrimental to object-level tasks, fragmenting coherent objects into many small clusters. We set $\tau = 0.6$ across tasks, providing a sensible trade-off between part sensitivity and semantic coherence.

Table 3. **Effect of clustering and aggregation.** We compare our approach with tuned baselines on COCO-20^i and PASCAL-Part (mIoU %, ↑). All thresholds (0.55) and clustering granularities (τ=0.5, τ=0.6) are tuned to optimal values for fair comparison.

Variant	COCO	PASCAL-Part
No clustering		
Thresholding sim. map @ 0.55 (Eq. 2)	44.2	35.4
Clustering w/o aggregation		
Coarse clustering ($\tau = 0.5$)	50.6	31.1
Fine clustering ($\tau = 0.6$)	42.8	36.2
Ours: *Clustering w/ aggregation* ($\tau = 0.6$)		
w/ cross similarity	54.6	48.5
w/ self + cross similarity (Eq. 13)	**57.6**	**50.5**

Clustering and aggregation. We lastly analyze the role of clustering and aggregation in Tab. 3. Thresholding the similarity map (Eq. 2) provides a coarse baseline, yielding 44.2 % and 35.4 % mIoU on COCO and PASCAL-Part. Introducing clustering without aggregation, where a single cluster is selected, improves results but requires tuning the granularity parameter τ separately for each task: lower values ($\tau = 0.5$) better capture full objects, while higher ones ($\tau = 0.6$) suit part-level structures. Even tuned independently, this variant still underperforms. Our solution fixes an intermediate τ and resolves this trade-off through a principled aggregation strategy (*cf*. Sec. 3.4). Aggregating by cross-image similarity already improves segmentation results, achieving 54.6 % and 48.5 %, while jointly leveraging cross- and self-similarity yields the best results, reaching 57.6 % and 50.5 % mIoU on COCO and PASCAL-Part.

5. Conclusion

In this work, we introduced INSID3, a training-free framework for in-context segmentation built solely on DINOv3. By leveraging the dual nature of DINOv3 features, *i.e.* semantic alignment and spatial coherence, INSID3 performs correspondence estimation and segmentation within a single backbone. Despite its simplicity, it exhibits strong generalization across one-shot *semantic*, *part*, and *personalized segmentation*, outperforming both fine-tuned and training-free approaches. Overall, this suggests that segmentation can emerge naturally from self-supervised dense representations. While existing methods rely on either fine-tuning or training-free pipelines grounded in mask-supervised pretraining, INSID3 remains fully *unsupervised*, relying solely on the in-context example for guidance. This suggests that reducing supervision may foster more robust and transferable representations, marking a concrete step toward more scalable and general-purpose visual understanding.

Acknowledgments. Claudia Cuttano was supported by the Sustainable Mobility Center (CNMS), which received funding from the European Union Next Generation EU (Piano Nazionale di Ripresa e Resilienza (PNRR), Missione 4 Componente 2 Investimento 1.4 "Potenziamento strutture di ricerca e creazione di 'campioni nazionali di R&S' su alcune Key Enabling Technologies") with grant agreement no. CN_00000023. Christoph Reich is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the German Federal Ministry of Education and Research. Stefan Roth has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008). Further, he was supported by the DFG under Germany's Excellence Strategy (EXC-3057/1 "Reasonable Artificial Intelligence", Project No. 533677015) and by the LOEWE initiative (Hesse, Germany) within the emergenCITY center [LOEWE/1/12/519/03/05.001(0016)/72]. Daniel Cremers has received funding by the European Research Council (ERC) Advanced Grant SIMULACRON (grant agreement No. 884679). We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources. We also acknowledge the support of the European Laboratory for Learning and Intelligent Systems (ELLIS). Finally, we thank Barış Zöngür for the insightful feedback.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. VICRegL: Self-supervised learning of local visual features. In *NeurIPS*, pages 8799–8810, 2022. 3

[2] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception Encoder: The best visual embeddings are not at the output of the network. In *NeurIPS*, 2025. 3

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, pages 1877–1901, 2020. 1, 2

[4] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P. Musco, Rahul K. Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J. McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Trans. Med. Imaging*, 33(2):577–590, 2013. 6

[5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, volume 14, pages 132–149, 2018. 3

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, pages 9912–9924, 2020. 3, 5

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2, 3

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 3

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, et al. PaLM: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(240):1–113, 2023. 1, 2

[10] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv:1902.03368 [cs.CV]*, 2019. 6

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 1

[12] Claudia Cuttano, Gabriele Trivigno, Giuseppe Averta, and Carlo Masone. SANSA: Unleashing the hidden semantics in SAM2 for few-shot segmentation. In *NeurIPS*, 2025. 2

[13] Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and predict latents patches for improved masked image modeling. *Trans. Mach. Learn. Res.*, 2025. 3

[14] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. 3

[15] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? In *CVPR*, pages 5414–5423, 2021. 3

[16] Miguel Espinosa, Chenhongyi Yang, Linus Ericsson, Steven McDonagh, and Elliot J. Crowley. No time to train! Training-free reference-based instance segmentation. *arXiv:2507.01300 [cs.CV]*, 2025. 3

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996. 5

[18] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.*, 32(11):1231–1237, 2013. 1

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Munos Rémi, and Valko Michal. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020. 3

[20] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):9052–9071, 2024. 3

[21] Oliver Hahn, Christoph Reich, Nikita Araslanov, Daniel Cremers, Christian Rupprecht, and Stefan Roth. Scene-centric unsupervised panoptic segmentation. In *CVPR*, pages 24485–24495, 2025. 6

[22] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *ICLR*, 2022. 3, 5, 6

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 3

[24] Greg Heinrich, Mike Ranzinger, Hongxu, Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. RADIOv2.5: Improved baselines for agglomerative vision foundation models. In *CVPR*, pages 22487–22497, 2025. 3

[25] Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron Van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *ICCV*, pages 10086–10096, 2021. 3

[26] Olivier J. Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. In *ECCV*, volume 27, pages 123–143, 2022. 3

[27] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4D convolutional Swin transformer for few-shot segmentation. In *ECCV*, pages 108–126, 2022. 2

[28] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *IROS*, pages 1769–1776, 2020. 6

[29] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. In *NeurIPS*, pages 19545–19560, 2020. 3

[30] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K. Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE Trans. Med. Imaging*, 33(2):233–245, 2013. 6

[31] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Found. Trends Comput. Graph. Vis.*, 12(1–3):1–308, 2020. 1

[32] Aleksandar Jevtić, Christoph Reich, Felix Wimbauer, Oliver Hahn, Christian Rupprecht, Stefan Roth, and Daniel Cremers. Feed-forward SceneDINO for unsupervised semantic scene completion. In *ICCV*, pages 6784–6796, 2025. 3

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *ICCV*, pages 4015–4026, 2023. 2, 3, 7

[34] Tae-young Ko and Seung-ho Lee. Novel method of semantic segmentation applicable to augmented reality. *Sensors*, 20 (6):1737, 2020. 1

[35] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 3

[36] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*, pages 109–117, 2011. 6

[37] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *CVPR*, pages 8057–8067, 2022. 2

[38] Yang Liu, Chenchen Jing, Hengtao Li, Muzhi Zhu, Hao Chen, Xinlong Wang, and Chunhua Shen. A simple image segmentation framework via in-context examples. In *NeurIPS*, pages 25095–25119, 2024. 1, 6

[39] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *ICLR*, 2024. 1, 2, 3, 4, 6

[40] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *CVPR*, pages 8364–8375, 2022. 5, 6

[41] Lingchen Meng, Shiyi Lan, Hengduo Li, Jose M. Alvarez, Zuxuan Wu, and Yu-Gang Jiang. SegIC: Unleashing the emergent correspondence for in-context segmentation. In *ECCV*, volume 38, pages 203–220, 2024. 1, 3, 6, 7

[42] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. SPair-71k: A large-scale benchmark for semantic correspondence. *arXiv:1908.10543 [cs.CV]*, 2019. 7

[43] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv:1109.2378 [stat.ML]*, 2011. 5, 6

[44] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *ICCV*, pages 622–631, 2019. 6

[45] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, volume 6, pages 69–84, 2016. 3

[46] Pedro O. Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C. Courville. Unsupervised learning of dense visual representations. In *NeurIPS*, pages 4489–4500, 2020. 3

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 1, 2, 3, 5

[48] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, pages 27730–27744, 2023. 1, 2

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3

[50] Miloš Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.*, 11(86):2487–2531, 2010. 5

[51] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, pages 12490–12500, 2024. 3

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 3

[53] Mohammadreza Salehi, Efstratios Gavves, Cees G.M. Snoek, and Yuki M. Asano. Time does tell: Self-supervised time-tuning of dense image representations. In *ICCV*, pages 16536–16547, 2023. 3

[54] Dvir Samuel, Rami Ben-Ari, Matan Levy, Nir Darshan, and Gal Chechik. Where's Waldo: Diffusion features for personalized segmentation and retrieval. In *NeurIPS*, pages 128160–128181, 2024. 6

[55] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.*, 42(3), 2017. 5

[56] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv:2508.10104 [cs.CV]*, 2025. 1, 2, 3, 4, 5, 6

[57] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *NeurIPS*, pages 1363–1389, 2023. 1, 7, 8

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971 [cs.CV]*, 2023. 1, 2

[59] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data*, 5(1):1–9, 2018. 6

[60] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *ICCV*, pages 10052–10062, 2021. 6

[61] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. PaNet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, pages 9197–9206, 2019. 2

[62] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K. Nandi. Medical image segmentation using deep learning: A survey. *IET Image Process.*, 16(5): 1243–1267, 2022. 1

[63] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, pages 6830–6839, 2023. 2, 6

[64] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards segmenting everything in context. In *ICCV*, pages 1130–1140, 2023. 2, 6

[65] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. In *ECCV*, volume 61, pages 320–337, 2024. 3

[66] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *CVPR*, pages 3987–3996, 2021. 3

[67] Junwei Yang, Ke Zhang, Zhaolin Cui, Jinming Su, Junfeng Luo, and Xiaolin Wei. InsCon: Instance consistency feature representation via self-supervised learning. *arXiv:2203.07688 [cs.CV]*, 2022. 3

[68] Xiwen Yao, Qinglong Cao, Xiaoxu Feng, Gong Cheng, and Junwei Han. Scale-aware detailed matching for few-shot aerial image semantic segmentation. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–11, 2021. 6

[69] Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. In *NeurIPS*, pages 33232–33261, 2024. 2, 3, 6, 7

[70] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable Diffusion complements DINO for zero-shot semantic correspondence. In *NeurIPS*, pages 45533–45547, 2023. 1, 7, 8

[71] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *CVPR*, pages 3076–3085, 2024. 7

[72] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. In *ICLR*, 2024. 1, 2, 6

[73] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *ICLR*, 2022. 3

[74] Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. In *NeurIPS*, pages 42672–42695, 2024. 1, 3, 6