Proceedings of 11th Workshop on Computer Vision in Plant Phenotyping and Agriculture, European Conference on Computer Vision (ECCV), September 8 – 12th, 2026, Malmö, Sweden

Self-Supervised Adaptation of Dense DINOv3 Features for Improved In-Context Vegetation Segmentation

Aleksandar Stefanov[1], Rajitha de Silva[2], Mamatha Thota[2], Riccardo Polvara[2], Marko Munih[1], and Matjaž Mihelj[1]

[1] University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia
{aleksandar.stefanov, marko.munih, matjaz.mihelj}@fe.uni-lj.si
[2] Lincoln Center for Autonomous Systems (L-CAS) & Applied Vision and Artificial Intelligence Lab (AVAIL), School of Engineering and Physical Sciences, University of Lincoln, UK
{mthota, odesilva, rpolvara}@lincoln.ac.uk

Abstract. Dense vegetation segmentation must transfer across fields, crops, viewpoints, and sensing platforms, but dense annotation for every condition is costly. In-context segmentation instead transfers a reference vegetation mask without task-specific segmentation training, placing strong demands on dense features: they must retain local detail for intra-crop transfer while tolerating cross-crop and cross-source appearance changes. We study this trade-off in DINOv3 features for agricultural imagery. DINOv3-L provides strong local specificity but can be overly appearance-specific across crops, whereas DINOv3-S yields a more tolerant similarity structure. We therefore train a lightweight residual adapter on frozen DINOv3-L features, using frozen DINOv3-S features to filter potential false negatives. On the Agroscapes dataset, within a fixed in-context segmentation pipeline, the adapted representation improves transfer within and across crops and between aerial and ground-based sources.

Keywords: DINOv3 · Self-Supervised Learning · Dense Visual Representations · In-Context Segmentation · Domain Adaptation

1 Introduction

Dense vegetation segmentation supports crop monitoring, phenotyping, navigation, and robotic intervention [14, 20]. Its transfer is challenging because crop type, growth stage, viewpoint, illumination, acquisition height, and sensing platform all change vegetation appearance [16]; collecting dense labels for every condition is therefore difficult to scale.

In-context segmentation transfers a reference annotation to a target image without task-specific training. Recent methods use dense correspondence, segmentation priors, or both to support semantic, part-level, and personalised transfer [8, 15, 17, 25, 29, 30]. Their performance depends on a feature space that is

locally discriminative for same-crop segmentation while remaining tolerant of different vegetation appearances.

Self-supervised foundation models are effective dense feature extractors for correspondence-driven segmentation. DINOv3 [21] is particularly suitable because its patch-level features support matching and region grouping. In our Agroscapes comparisons, DINOv3-L preserves fine local structure but can be appearance-specific across crops, whereas DINOv3-S provides a more tolerant similarity signal.

We propose teacher-filtered DINOv3 feature adaptation for in-context vegetation segmentation. Both encoders remain frozen: a lightweight residual adapter transforms DINOv3-L features, while DINOv3-S estimates patch similarity to exclude potentially related cross-image patches from the contrastive negative set. An anchoring term keeps adapted features aligned with DINOv3-L features.

We evaluate on Agroscapes from Agronav [19] in a fixed, INSID3-style [8] training-free pipeline, varying only the dense representation. Across mixed-source and aerial-to-ground splits, the adapted features retain strong intra-crop performance while improving cross-crop and cross-source transfer.

Contributions. Our contributions are:

- We analyze how DINOv3 large and small variants behave under agricultural domain shifts, showing a trade-off between fine-grained local specificity and broader cross-crop semantic tolerance.
- We propose a lightweight self-supervised adaptation method for DINOv3 large-variant features, using DINOv3 small-variant teacher filtering to remove potential false negatives and DINOv3 large-variant anchoring to preserve the original feature geometry.
- We evaluate the adapted representation in a fixed INSID3-style pipeline and show improved intra-crop, cross-crop, mixed-source, and aerial-to-ground vegetation segmentation transfer.

2 Related Work

Agricultural Semantic Segmentation and Domain Transfer. Agricultural segmentation supports monitoring, phenotyping, and robotic perception [14, 19], including navigation through vegetation and traversable-region estimation [1, 7, 22]. In Agronav-style deployments, a segmentation system must distinguish vegetation from soil and other background regions despite changes in crop morphology and image formation. Crop type, growth stage, illumination, viewpoint, acquisition height, field layout, and sensing platform consequently create substantial domain shifts. A representation that is locally precise for one field can therefore fail to associate semantically related vegetation in another field, especially when the canopy appearance or viewing geometry changes. Prior work addresses transfer across crops [6], agricultural fields and robots [16], crop–weed domains [27], and ground/UAV imagery [9], typically by adapting or training segmentation models. In contrast, we self-supervisedly adapt the dense representation and use masks only for downstream evaluation. This isolates whether

feature-space adaptation itself improves the correspondence required by an in-context segmenter.

Dense Feature Representations for In-Context Segmentation. In-context segmentation requires features that support cross-image correspondence and within-image grouping. The reference annotation is not sufficient by itself: its utility depends on whether foreground descriptors remain comparable to the corresponding target regions while remaining distinct from background descriptors. Foundation models provide such features through correspondence, segmentation priors, or dense self-supervision [8, 15, 17, 29]; INSID3 specifically uses DINOv3 features for training-free matching, clustering, and aggregation [8]. Self-supervision is also increasingly used in precision agriculture [2, 11, 13], and agricultural foundation models such as Agri-FM+ [3] and SPROUT [28] reduce the gap between generic and agricultural representations. These approaches motivate studying representation transfer separately from a task-specific decoder. Rather than pretraining a new model or modifying the downstream pipeline, we adapt frozen DINOv3-L features and use DINOv3-S only for false-negative filtering.

Dense Vision Foundation Models under Domain Shift. DINO patch features encode dense semantic descriptors for correspondence [4], and DINOv2 [18] demonstrated the transferability obtained by scaling self-supervision. DINOv3 [21] extends this family with high-quality dense representations, while recent work has adapted DINO descriptors for dense prediction [23]. For dense in-context transfer, however, a feature is useful only if its local detail does not make it overly sensitive to crop-specific texture, row structure, or viewpoint. We examine this trade-off empirically for DINOv3-S and DINOv3-L on agricultural domain shift; it motivates the complementary large-student/small-teacher roles in our adaptation objective rather than a general claim about model scale.

Parameter-Efficient Adaptation of Vision Foundation Models. Frozen vision foundation models can be adapted with lightweight modules rather than full fine-tuning [18, 21]. This is particularly attractive when the available agricultural data are unlabelled and the downstream method consumes a dense feature grid rather than task logits. DINOv3-FD [12] and adapter-based Grounding DINO tuning [26] show that such adaptation can specialise representations under domain shift or limited data. Teacher guidance is also relevant to contrastive learning because visually different but semantically related vegetation patches may otherwise be treated as negatives. DINOv3 further uses Gram anchoring during pretraining [21]. False-negative filtering has also been used in self-supervised domain adaptation to avoid repelling semantically similar samples [24]. Our method combines lightweight adaptation with DINOv3-S-guided negative filtering and an anchor to the original DINOv3-L features. The filter and anchor serve distinct purposes: the former changes which cross-image relations are repelled, whereas the latter constrains the adapted features to remain close to the pretrained representation.

Evaluation of Foundation Model Representations for Cross-Domain Vegetation Segmentation. INSID3 [8] shows that one frozen DINOv3 back-

bone can support training-free in-context segmentation, while controlled geospatial studies find that frozen features can match fine-tuning for agricultural field segmentation [5]. CrossEarth [10] and agricultural ground/UAV transfer studies [9] further highlight the importance of evaluation under domain shift. In this setting, reporting only an aggregate score can obscure whether a representation preserves same-crop matching while improving the more demanding cross-crop cases. We therefore report intra-crop and cross-crop transfers under both mixed-source and aerial-to-ground training regimes. Unlike work that changes the foundation model or downstream architecture, we vary only the dense representation within a fixed INSID3-style pipeline, isolating its effect on the four transfer settings.

3 Methodology

3.1 Problem Overview and Method Motivation

Given a reference image I_r with binary vegetation mask M_r and a target image I_t, we segment the corresponding vegetation in I_t, including cases with different crop types or appearances. The mask is available only to the fixed downstream in-context procedure; adapter training uses no vegetation annotations. Thus, the objective is not to learn a segmentation classifier, but to make the feature representation more transferable for reference-to-target matching.

Let $E(\cdot)$ denote a frozen dense visual encoder that maps an image I to a spatial feature map,

$$F = E(I) \in \mathbb{R}^{C\times H\times W}. \tag{1}$$

Rather than training a task-specific segmentation network, we learn a lightweight adapter A_θ that transforms the frozen features into an adapted representation,

$$Z = A_\theta(F). \tag{2}$$

The self-supervised adapter is designed to retain local discriminability while improving correspondence across crop types and acquisition conditions; it is evaluated in a fixed in-context segmentation pipeline. This formulation separates representation learning from the downstream clustering and aggregation choices, so any change in segmentation quality can be attributed to the features supplied to the same procedure.

3.2 Dense DINOv3 Features for Agricultural Transfer

To analyse dense DINOv3 features under agricultural domain shift, let $E_L(\cdot)$ and $E_S(\cdot)$ denote frozen DINOv3-L and DINOv3-S encoders that produce

$$F_L = E_L(I) \in \mathbb{R}^{C_L\times H\times W}, \qquad F_S = E_S(I) \in \mathbb{R}^{C_S\times H\times W}. \tag{3}$$

For a patch location p, we denote the corresponding normalized features as

$$f_L(I,p) = \frac{F_L(p)}{\|F_L(p)\|_2}, \qquad f_S(I,p) = \frac{F_S(p)}{\|F_S(p)\|_2}. \tag{4}$$

The cosine similarity between patches p_i and p_j is then

$$s_L(p_i, p_j) = f_L(I_i, p_i)^\top f_L(I_j, p_j), \qquad s_S(p_i, p_j) = f_S(I_i, p_i)^\top f_S(I_j, p_j). \quad (5)$$

We use these similarities to characterise transfer rather than to supervise a segmentation model. For two views of the same crop, high similarity at corresponding vegetation locations is desirable, but cross-crop transfer additionally requires features to remain comparable when leaf shape, canopy density, colour, and background context differ. The qualitative comparison in Fig. 1 exposes a practical scale trade-off: fine-grained descriptors can improve local boundary discrimination while assigning low affinity to vegetation whose appearance changes substantially. Conversely, a coarser descriptor can be informative about broad semantic relatedness even when it does not preserve the local detail required by clustering.

Fig. 1 illustrates that DINOv3-L preserves fine local structure but becomes more appearance-specific across crops, whereas DINOv3-S produces less detailed yet more tolerant features. The large variant separates nearby local structures that are useful for delineating vegetation, but its cross-image similarities can decrease when crop texture and viewing conditions change. The small variant provides a coarser semantic signal that is less suitable as the final dense descriptor but useful for identifying pairs that should not be strongly repelled.

This complementary behaviour motivates an asymmetric design: DINOv3-L is adapted for its local specificity, while DINOv3-S only identifies potentially related cross-image patch pairs. The adapted representation is computed solely from DINOv3-L features,

$$Z = A_\theta(F_L), \quad (6)$$

and DINOv3-S guides the false-negative filtering in the contrastive objective. Consequently, DINOv3-S neither provides input channels to the adapter nor changes the representation used at inference; its role is restricted to defining a more reliable set of contrastive negatives during training.

3.3 Teacher-Filtered Patch-Level Contrastive Adaptation

Fig. 2 summarises training. Both DINOv3 encoders remain frozen; only A_θ is optimised. DINOv3-L provides the adapted representation and DINOv3-S filters potential false negatives. This keeps the trainable component small and avoids changing either foundation-model feature extractor while exploiting their complementary similarity behaviour.

For each training image I_i, the frozen encoders produce

$$F_L^i = E_L(I_i), \qquad F_S^i = E_S(I_i). \quad (7)$$

The same random spatial crop and optional horizontal flip are applied to both feature maps,

$$\bar{F}_L^i, \bar{F}_S^i = \mathcal{C}(F_L^i, F_S^i), \quad (8)$$

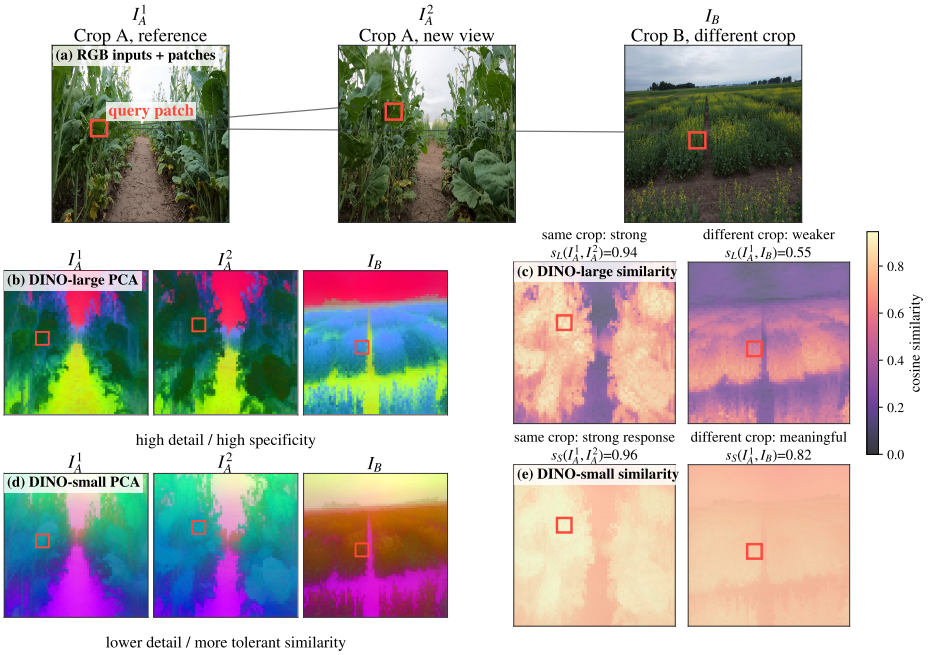


Fig. 1: Comparison of dense DINOv3 features for agricultural transfer. The top row shows two views of the same crop and one image of a different crop. DINOv3-L preserves finer local structure but produces lower cross-crop similarity, whereas DINOv3-S yields a more tolerant response across vegetation appearances.

preserving correspondence between the DINOv3-L and DINOv3-S grids. Applying the transform jointly is important because the teacher similarity at a patch location is used to decide whether the corresponding large-model candidate may be a negative. Independent spatial transforms would invalidate this teacher–student alignment.

Two normalised views are generated from the cropped DINOv3-L features by adding independent Gaussian noise,

$$\tilde{F}^i_{L,a} = \mathrm{norm}\big(\bar{F}^i_L + \epsilon_a\big)\,, \qquad \tilde{F}^i_{L,b} = \mathrm{norm}\big(\bar{F}^i_L + \epsilon_b\big)\,, \tag{9}$$

where ϵ_a and ϵ_b are independently sampled noise tensors. The perturbations create two views of the same underlying patch grid without requiring a separate image augmentation pipeline. The shared adapter processes both views,

$$Z^i_a = A_\theta(\tilde{F}^i_{L,a}), \qquad Z^i_b = A_\theta(\tilde{F}^i_{L,b}). \tag{10}$$

Let $z_a^{i,p}$ and $z_b^{i,p}$ denote normalised adapted features at patch p. Corresponding locations form positives, so the objective explicitly preserves each patch identity across the two perturbations. Candidate negatives are drawn from other batch images rather than from different locations in the same image; they are then filtered with DINOv3-S teacher features. This focuses the contrastive pres-

sure on cross-image separation while avoiding a forced distinction between neighbouring patches that may belong to the same vegetation region.

For anchor patch p in I_i and candidate patch q in I_j, $i \neq j$, teacher similarity is

$$s_S(i,p,j,q) = \frac{\bar{F}_S^i(p)^\top \bar{F}_S^j(q)}{\|\bar{F}_S^i(p)\|_2 \, \|\bar{F}_S^j(q)\|_2}. \tag{11}$$

The filtered negative set is

$$\mathcal{N}_{i,p} = \{(j,q) \mid j \neq i,\ s_S(i,p,j,q) < \tau_S\}, \tag{12}$$

where τ_S is the teacher threshold; pairs above it are excluded as potentially semantically related vegetation. The rule is deliberately conservative: a high DINOv3-S similarity does not assign a positive label, but prevents the loss from treating that pair as certainly negative. This addresses the false-negative mechanism without introducing labels or requiring explicit cross-crop correspondences.

Filtering is applied independently for each anchor patch and each candidate from a different batch image. It therefore retains visually or semantically dissimilar patches as contrastive negatives, while discarding only those for which the smaller frozen model supplies evidence of relatedness. This distinction is important in agricultural imagery: different images can contain repeated vegetation patterns, yet not all cross-image patch pairs should collapse into one representation. The thresholded teacher signal supplies a lightweight uncertainty test rather than a class assignment.

The contrastive loss for anchor $z_a^{i,p}$ is

$$\ell_{i,p} = -\log \frac{\exp\Big(\mathrm{sim}(z_a^{i,p}, z_b^{i,p})/\tau_c\Big)}{\exp\Big(\mathrm{sim}(z_a^{i,p}, z_b^{i,p})/\tau_c\Big) + \sum\limits_{(j,q)\in\mathcal{N}_{i,p}} \exp\Big(\mathrm{sim}(z_a^{i,p}, z_b^{j,q})/\tau_c\Big)}, \tag{13}$$

where $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity and τ_c is the contrastive temperature. Each retained candidate contributes to the denominator, so the loss attracts the same-location positive while repelling only candidates supported as dissimilar by the frozen teacher. Averaging over images and patches gives

$$\mathcal{L}_{pc} = \frac{1}{|\mathcal{B}|\,|\mathcal{P}|} \sum_{i\in\mathcal{B}} \sum_{p\in\mathcal{P}} \ell_{i,p}. \tag{14}$$

This objective promotes perturbation-invariant features without separating potentially related cross-image vegetation patches. It does not assert that all teacher-similar regions share the same class; instead, it removes the strongest contradictory supervision signal for pairs whose semantic relationship is uncertain under crop and platform shift.

At inference, this training construction is not retained as an additional processing stage. A target image is encoded once by frozen DINOv3-L and passed through the learned adapter; the fixed INSID3 pipeline then consumes the adapted grid in exactly the same way as the original DINOv3-L grid. The teacher is consequently a training-time source of negative-set supervision, not an ensemble component or an additional inference dependency.

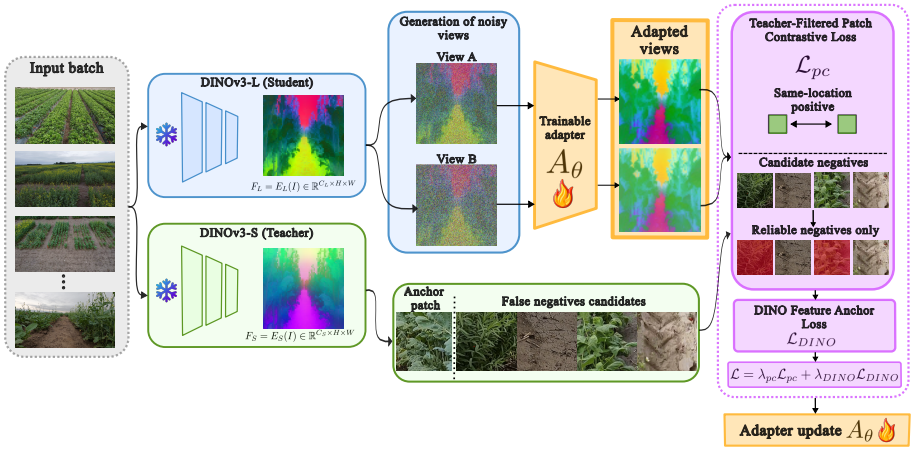


Fig. 2: Overview of the teacher-filtered DINOv3 adaptation pipeline. Frozen DINOv3-L features are perturbed to form two views and passed through the shared adapter. Same-location patches form positive pairs, while cross-image patches form candidate negatives. Frozen DINOv3-S features remove candidate negatives with high teacher similarity. The adapter is trained using the filtered patch-level contrastive loss and the DINOv3-L feature anchoring loss.

3.4 Adapter Architecture

The lightweight residual adapter A_θ preserves the DINOv3-L spatial grid, mapping $F_L \in \mathbb{R}^{C_L \times H \times W}$ to $Z \in \mathbb{R}^{C_Z \times H \times W}$. Here $C_Z = C_L = 1024$, so the adapted representation retains the original resolution and dimensionality. Maintaining this interface allows the adapter output to replace the frozen DINOv3-L features directly in INSID3, without changing its matching, clustering, or mask-aggregation stages.

Because $C_Z = C_L$, the input projection is an identity. A gated multi-head correction block produces $\Delta_\theta(F_L)$, yielding the residual update

$$Z' = F_L + \alpha\, \Delta_\theta(F_L), \tag{15}$$

where α is a residual scaling factor. The residual form makes the adaptation an explicit correction to the pretrained feature map rather than a complete re-encoding of it.

A small output MLP with 1×1 convolutions then refines and channel-normalises the residual feature map:

$$Z = \mathrm{norm}\left(Z' + \phi_{\mathrm{out}}(Z')\right). \tag{16}$$

The resulting dense, L_2-normalised representation is used directly by the downstream pipeline. Normalisation maintains the cosine-similarity interpretation used by the contrastive objective and dense matching procedure. The adapter has 8 attention heads, 512 hidden channels, and 1024 output channels.

The architecture is intentionally restricted to a residual correction block and a small output MLP. It can adjust the similarity structure responsible for

reference-to-target association while retaining the spatial sampling and channel dimensionality expected by the downstream system. This design also makes the adaptation compatible with the cached-feature training protocol: the adapter operates on frozen DINOv3-L feature maps rather than requiring repeated end-to-end optimisation of the encoder.

3.5 Feature-Space Regularization

The patch-level contrastive objective alone may over-specialise the representation to sampled training pairs. We therefore anchor adapted features to frozen DINOv3-L features, retaining useful pretrained local structure. Unlike a full geometry-preservation constraint, the anchor is a pointwise similarity regulariser: it preserves correspondence to the original feature at each patch while leaving the adapter free to change cross-image relations needed by the filtered contrastive objective.

For a normalized adapted patch feature $z_a^{i,p}$ and the corresponding normalized DINOv3-L feature $f_L^{i,p}$, we define the feature anchoring loss as

$$\mathcal{L}_{DINO} = 1 - \frac{1}{|\mathcal{B}|\,|\mathcal{P}|} \sum_{i \in \mathcal{B}} \sum_{p \in \mathcal{P}} \left(z_a^{i,p}\right)^\top f_L^{i,p}. \tag{17}$$

This term penalises excessive deviation from DINOv3-L while allowing adaptation through the contrastive objective. It therefore counterbalances the pressure to specialise on the self-supervised training set and helps retain the locally discriminative information observed for DINOv3-L.

The final self-supervised training objective is

$$\mathcal{L} = \lambda_{pc}\mathcal{L}_{pc} + \lambda_{DINO}\mathcal{L}_{DINO}, \tag{18}$$

where λ_{pc} and λ_{DINO} weight the contrastive and anchoring terms. Both are 1.0 in our experiments. Together, the objectives promote stable patch descriptors, prevent uncertain cross-image vegetation pairs from being repelled, and retain alignment with the pretrained DINOv3-L representation.

4 Experiments and Results

We evaluate feature adaptation for in-context vegetation segmentation with INSID3 [8] as a fixed downstream protocol. INSID3 transfers a reference mask through dense matching, agglomerative clustering, seed-cluster selection, and cluster aggregation. We fix reference-target pairs, masks, clustering settings, and metrics, varying only the dense representation. The resulting mIoU, precision, and recall therefore probe how well each representation supports the same reference-conditioned segmentation process rather than the effect of a different decoder or task-specific training recipe. The conclusions are limited to the six Agroscapes sequences and transfer settings reported here.

Implementation details. We train the adapter self-supervisedly on cached DINOv3-L features, using frozen DINOv3-S features only for teacher-based negative filtering. Following INSID3 [8], all DINOv3-based representations are debiased before training and evaluation using rank $s = 500$. The adapter is optimized for 50 epochs with AdamW, a constant learning rate of 5×10^{-5}, weight decay 10^{-4} and $\tau_S = 0.7$.

Input images are resized to 1024×1024. During training, we sample random 24×24 feature crops, apply horizontal flipping with probability 0.5, and generate two views using Gaussian feature noise with standard deviation 0.03. Up to 256 patch locations are sampled uniformly without replacement from each crop, and the patch-contrastive temperature is $\tau_c = 0.1$. Batch sizes are 16 for mixed-source training and 4 for aerial-to-ground training, whose training set is smaller after excluding aerial canola to avoid crop-level leakage. Main results use a fixed-seed checkpoint, while seed variability is assessed using training seeds $\{11, 22, 33\}$. The trained adapter is then frozen and used only as a dense feature extractor. All experiments are performed on one NVIDIA RTX PRO 6000 Blackwell GPU with 96 GB of GDDR7 ECC memory.

Debiasing, reference–target pairs, and clustering parameters are held fixed wherever DINOv3-based representations are compared. Thus, the only difference between frozen DINOv3-L and our representation is the learned residual transformation of the same debiased features.

4.1 Dataset and Evaluation Setup

We evaluate on the Agroscapes dataset from Agronav [19], which provides pixel annotations across crop types, viewpoints, acquisition heights, and aerial and ground-based platforms. We use six sequences (Table 1) and convert vegetation annotations to binary masks. The selected sequences include strawberry, canola, bean, and flax, and combine aerial/drone acquisition with ground-robot imagery. This variation makes the evaluation sensitive to both semantic appearance changes between crops and geometric changes induced by platform and camera height.

Mixed-source split. The adapter is trained on `DJI_048`, `DJI_175`, `DJI_234`, and `GH_10005`, and evaluated on held-out `DJI_280` and `GH_10041`. This setting evaluates transfer to unseen crop types and acquisition conditions with aerial and ground-based self-supervised training data. It tests whether exposure to heterogeneous unlabelled imagery can improve matching on the held-out bean and flax sequences without using their masks during adaptation.

Aerial-to-ground split. The adapter is trained only on aerial data and evaluated on `GH_10005` and `GH_10041`. Aerial canola sequences `DJI_175` and `DJI_234` are excluded to avoid crop-level leakage because evaluation includes ground-based canola. This split assesses transfer across crop appearance and platform while making the canola result a platform-transfer test rather than a same-crop adaptation result.

For both splits, intra-crop transfer uses the same or visually similar crops, whereas cross-crop transfer uses different crop types.

Table 1: Agroscapes sequences used in the experiments, covering multiple crop types and acquisition settings.

Sequence	Crop	Setting	Images
DJI_048	Strawberry	Aerial / drone	51
DJI_175	Canola	Aerial / drone	160
DJI_234	Canola	Aerial / drone, diff. height	77
DJI_280	Bean	Aerial / drone	119
GH_10005	Canola	Ground robot	435
GH_10041	Flax	Ground robot	321

Table 2: Mixed-source split results with methods as rows and transfer setups as grouped columns. Each setup reports mean IoU (mIoU), precision, and recall; all results are stated in %. Best results within each setup are bold and second-best values are underlined.

	Flax → Flax			Bean → Flax			Flax → Bean			Bean → Bean		
Method	mIoU	Prec.	Rec.	mIoU	Prec.	Rec.	mIoU	Prec.	Rec.	mIoU	Prec.	Rec.
DINOv3-L	<u>88.2</u>	**98.6**	89.4	3.8	57.9	3.8	50.7	<u>85.6</u>	54.0	<u>74.1</u>	**91.1**	80.0
SPROUT-dec.	60.9	62.5	**96.5**	<u>61.0</u>	62.6	**96.6**	49.7	51.7	<u>93.0</u>	50.7	52.1	**95.3**
SPROUT-bott.	60.2	62.6	<u>95.0</u>	60.5	<u>62.7</u>	<u>95.4</u>	<u>56.9</u>	59.5	**93.5**	57.4	60.1	<u>93.5</u>
Ours	**91.6**	<u>97.3</u>	94.0	**90.6**	**97.1**	92.9	**76.0**	**91.1**	82.1	**75.9**	**91.1**	82.2

4.2 Main results

Baselines. We compare against frozen DINOv3-L in the original INSID3 configuration and two SPROUT [28] representations: bottleneck features before decoding and spatially refined decoder features. SPROUT feature maps are resized to the DINOv3 grid and used in the unchanged INSID3 pipeline. Our method applies A_θ to frozen, debiased DINOv3-L features. The comparison consequently contrasts representations at the same downstream interface, rather than comparing segmentation heads trained with different supervision. Unless stated otherwise, all methods use $\tau = 0.6$ and $\alpha = 0.2$; SPROUT-bottleneck uses $\tau = 0.75$ for aerial-to-ground evaluation because the default produced overly coarse clusters.

Intra-crop and Cross-crop Segmentation on Mixed-source Split. Table 2 reports the mixed-source split. The adapted representation achieves the best mIoU in all four transfer cases, improving DINOv3-L by +3.4 and +1.8 points for Flax → Flax and Bean → Bean, respectively. These within-crop gains show that the adapter does not trade away the strong same-crop behaviour of the large model. Gains are larger for Bean → Flax (+86.8 points) and Flax → Bean (+25.3 points), where frozen DINOv3-L has low recall or reduced overlap. In the Bean → Flax case, mIoU increases from 3.8% to 90.6%, accompanied by recall increasing from 3.8% to 92.86%; this is consistent with improved cross-crop correspondence rather than a precision-only effect. Compared with SPROUT, the proposed features generally retain higher precision while improving recall in difficult transfers. Qualitative results appear in Fig. 3.

Table 3: Aerial-to-ground-source split results with methods as rows and transfer setups as grouped columns. Each setup reports mean IoU (mIoU), precision, and recall; all results are stated in %. Best results within each setup are bold and second-best values are underlined.

	Canola → Canola			Flax → Canola			Canola → Flax			Flax → Flax		
Method	mIoU	Prec.	Rec.	mIoU	Prec.	Rec.	mIoU	Prec.	Rec.	mIoU	Prec.	Rec.
DINOv3-L	<u>81.4</u>	<u>98.4</u>	82.6	<u>70.6</u>	<u>98.5</u>	71.2	<u>79.1</u>	<u>98.0</u>	79.9	**88.2**	<u>98.6</u>	89.4
SPROUT-dec.	60.7	68.5	**85.9**	59.7	68.3	**84.7**	61.0	62.6	<u>96.6</u>	60.9	62.5	**96.5**
SPROUT-bott.	62.0	73.4	83.1	56.7	72.9	76.1	63.0	64.8	**96.8**	61.6	65.0	<u>93.9</u>
Ours	**83.4**	**98.9**	<u>84.6</u>	**81.0**	**99.1**	<u>81.1</u>	**85.5**	**99.0**	86.2	<u>86.0</u>	**99.1**	86.8

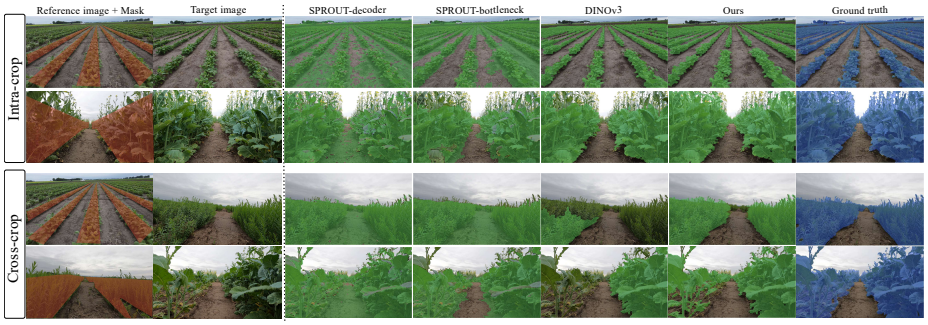


Fig. 3: Qualitative comparison of in-context vegetation segmentation for intra-crop and cross-crop transfer. The first and third rows show examples from the mixed-source split, while the second and fourth rows show examples from the aerial-to-ground split.

Intra-crop and Cross-crop Segmentation on Aerial-to-Ground-source Split. Table 3 reports aerial-to-ground transfer, where the adapter is trained only on aerial data and evaluated on ground-robot sequences. The proposed method improves DINOv3-L in three of four cases: by +2.0 points for Canola → Canola, +10.4 for Flax → Canola, and +6.4 for Canola → Flax. These gains are obtained despite the change in sensing platform and, for cross-crop cases, the change in vegetation appearance. It also achieves the highest precision in all four cases, whereas SPROUT generally attains higher recall but lower mIoU; the proposed representation therefore yields a more balanced foreground transfer in this fixed clustering pipeline. DINOv3-L remains stronger for Flax → Flax by 2.2 points, indicating that adaptation does not improve every already-strong same-crop pairing. Qualitative results appear in Fig. 3.

4.3 Robustness to the Clustering Threshold

The clustering threshold τ controls agglomerative-clustering granularity: lower values produce coarser regions and higher values finer partitions. We sweep τ for DINOv3-L and the adapted representation to assess dependence on this downstream setting.

Fig. 4 shows that the adapted representation is stronger across the tested thresholds, particularly for challenging cross-crop transfer, where DINOv3-L re-

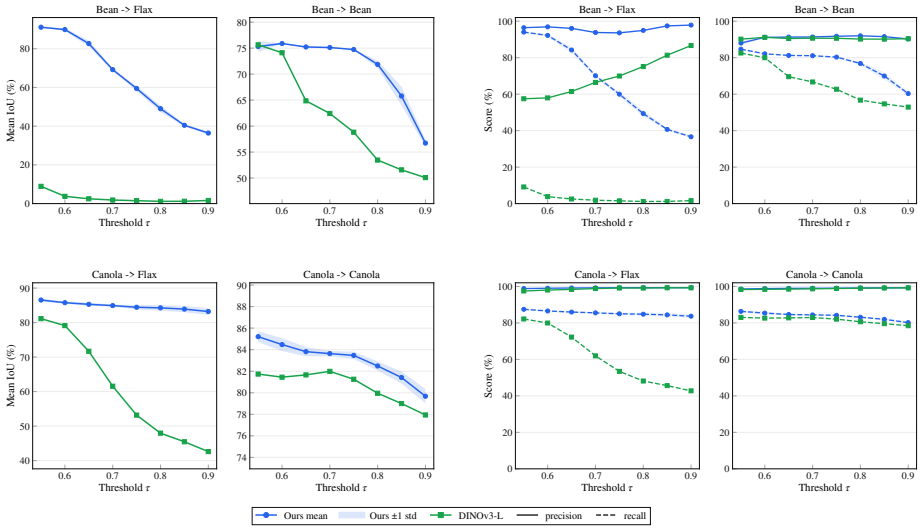


Fig. 4: Sensitivity to the agglomerative clustering threshold τ. The proposed adapted representation achieves higher and more stable mIoU than DINOv3-L across the tested intra-crop and cross-cro p transfer settings, particularly in the challenging cross-crop cases.

call and mIoU are low. The same trend is observed for Canola → Flax; intra-crop differences are smaller because DINOv3-L is already strong. Threshold sweeps are relevant here because clustering can otherwise confound a representation comparison: a gain observed only at one partition granularity could reflect threshold tuning rather than improved descriptors. In these evaluated settings, the ordering is not tied to a single τ.

4.4 Ablation Study

Table 4 evaluates both components on the mixed-source split. Removing either the DINOv3-L anchor or DINOv3-S teacher filter degrades performance, especially for cross-crop transfer. Removing the anchor reduces mIoU by 49.5 points for Flax → Bean and by 75.5 points for Bean → Flax. Removing teacher filtering reduces mIoU by 24.2 and 73.4 points in the same two directions, respectively. The complementary pattern is consistent with the method design. Without anchoring, contrastive adaptation can move the features too far from the locally discriminative large-model representation; without teacher filtering, semantically related vegetation patches can remain in the negative set. Removing both terms is harmful overall, while the full model is best in every setting.

4.5 Computational analysis

On the same 756-image evaluation set, the proposed representation requires 321.6 ms per image versus 307.5 ms for DINOv3-L INSID3: a 14.1 ms (4.6%) overhead.

Table 4: Component ablation results on the mixed-source split, measured by mean IoU (mIoU) at $\tau = 0.6$. All results are stated in %.

Variant	Flax → Flax	Flax → Bean	Bean → Flax	Bean → Bean
No DINOv3-L anchor	88.7	26.5	15.1	65.4
No DINOv3-S teacher filter	80.9	51.8	17.2	62.6
No DINOv3-L anchor, no DINOv3-S teacher filter	81.1	45.0	10.7	60.0
Full model	**91.6**	**76.0**	**90.6**	**75.9**

DINOv3 forward time is unchanged (170.2 vs. 170.0 ms); the adapter adds 12.1 ms, while clustering remains similar (134.7 vs. 131.5 ms). The adaptation cost is therefore incurred after the frozen encoder and does not require a second backbone at evaluation time. Thus, the method adds modest cost without changing the overall INSID3 profile.

5 Conclusion

We studied DINOv3 feature adaptation for in-context vegetation segmentation under agricultural domain shift. In the evaluated Agroscapes settings, DINOv3-L provides local detail but can be appearance-specific across crops, whereas DINOv3-S supplies a more tolerant similarity signal. Our teacher-filtered residual adapter uses frozen DINOv3-S features to suppress potential false negatives and anchors outputs to DINOv3-L features. The asymmetric design keeps the large model as the only inference backbone and uses the smaller model solely to prevent uncertain cross-image pairs from imposing contradictory contrastive supervision. Within the fixed in-context pipeline, the adapted representation improves the reported within-crop, cross-crop, aerial, and ground-based transfers without task-specific segmentation training. These results support lightweight feature-space adaptation as a way to improve the evaluated reference-conditioned vegetation-transfer settings while retaining the dense interface and local structure relied upon by the downstream pipeline.

Acknowledgements

The study was funded by the Slovenian Research and Innovation Agency (ARIS) (research programme code P2-0228) and as part of the project Robot4Plants (Contract No. 33126-1/2025/18), funded under the IRP38 intervention from the Common Agricultural Policy 2023–2027. The research was also co-funded by ARIS and the Ministry of Agriculture of the Republic of Slovenia within the project "Versatile AI-Driven Robotic System for Plant and Produce Management" (research project code L2-60143).

References

1. Aghi, D., Cerrato, S., Mazzia, V., Chiaberge, M.: Deep semantic segmentation at the edge for autonomous navigation in vineyard rows. In: 2021 IEEE/RSJ In-

ternational Conference on Intelligent Robots and Systems (IROS). pp. 3421–3428 (2021). https://doi.org/10.1109/IROS51168.2021.9635969
2. Al Mamun, A., Ahmedt-Aristizabal, D., Zhang, M., Hossen, M.I., Hayder, Z., Awrangjeb, M.: Plant disease detection using self-supervised learning: A systematic review. IEEE Access (2024)
3. Al Nahian, M.J., Ghosh, T., Sheikhi, F., Maleki, F.: Agri-fm+: A self-supervised foundation model for agricultural vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 5550–5562 (2025)
4. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. ECCVW What is Motion For? (2022), https://dino-vit-features.github.io/
5. Anonymous: Frozen features often match fine-tuning: A controlled evaluation of geospatial foundation models for agricultural field-extent segmentation (2026), https://openreview.net/forum?id=qRXVTe1yYp, under review at Transactions on Machine Learning Research
6. Bosilj, P., Aptoula, E., Duckett, T., Cielniak, G.: Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture. Journal of Field Robotics **37**(1), 7–19 (2020). https://doi.org/10.1002/rob.21869
7. Cao, M., Tang, F., Ji, P., Ma, F.: Improved real-time semantic segmentation network model for crop vision navigation line detection. Frontiers in Plant Science **13**, 898131 (2022). https://doi.org/10.3389/fpls.2022.898131
8. Cuttano, F., Gabriele, T., Christoph, R., Daniel, C., Carlo, M., Roth, S.: Insid3: Training-free in-context segmentation with dinov3. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
9. Gao, J., Liao, W., Nuyttens, D., Lootens, P., Alexandersson, E., Pieters, J.: Cross-domain transfer learning for weed segmentation and mapping in precision farming using ground and uav images. Expert Systems with Applications **246**, 122980 (2024). https://doi.org/10.1016/j.eswa.2023.122980
10. Gong, Z., Wei, Z., Wang, D., Ma, X., Chen, H., Jia, Y., Deng, Y., Ji, Z., Zhu, X., Yokoya, N., Zhang, J., Du, B., Zhang, L.: Crossearth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. arXiv preprint arXiv:2410.22629 (2024), https://arxiv.org/abs/2410.22629
11. Gustineli, M., Miyaguchi, A., Stalter, I.: Multi-label plant species classification with self-supervised vision transformers. arXiv preprint arXiv:2407.06298 (2024)
12. He, Z., Fu, Y., Jin, Y.: Incentivizing dinov3 adaptation for medical vision tasks via feature disentanglement. In: Proceedings of The 9th International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 315, pp. 2496–2513. PMLR (2026), https://proceedings.mlr.press/v315/he26a.html
13. Li, Y., Zhan, X., Liu, S., Lu, H., Jiang, R., Guo, W., Chapman, S., Ge, Y., Solan, B., Ding, Y., et al.: Self-supervised plant phenotyping by combining domain adaptation with 3d plant model simulations: Application to wheat leaf counting at seedling stage. Plant Phenomics **5**, 0041 (2023)
14. Li, Z., Guo, R., Li, M., Chen, Y., Li, G.: A review of computer vision technologies for plant phenotyping. Computers and Electronics in Agriculture **176**, 105672 (2020). https://doi.org/10.1016/j.compag.2020.105672

15. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. In: International Conference on Learning Representations (2024)
16. Magistri, F., Weyler, J., Gogoll, D., Lottes, P., Behley, J., Petrinic, N., Stachniss, C.: From one field to another – unsupervised domain adaptation for semantic segmentation in agricultural robotics. Computers and Electronics in Agriculture **212**, 108114 (2023). https://doi.org/10.1016/j.compag.2023.108114
17. Meng, L., Lan, S., Li, H., Alvarez, J.M., Wu, Z., Jiang, Y.G.: Segic: Unleashing the emergent correspondence for in-context segmentation. In: European Conference on Computer Vision. pp. 203–220 (2024)
18. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., 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., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023), https://arxiv.org/abs/2304.07193
19. Panda, S.K., Lee, Y., Khalid Jawed, M.: Agronav: Autonomous Navigation Framework for Agricultural Robots and Vehicles using Semantic Segmentation and Semantic Line Detection . In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 6272–6281 (2023)
20. de Silva, R., Cielniak, G., Wang, G., Gao, J.: Deep learning-based crop row detection for infield navigation of agri-robots. Journal of Field Robotics **41**(7), 2299–2321 (2024). https://doi.org/10.1002/rob.22238
21. 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. arXiv preprint arXiv:2508.10104 (2025), https://arxiv.org/abs/2508.10104
22. Song, Y., Xu, F., Yao, Q., Liu, J., Yang, S.: Navigation algorithm based on semantic segmentation in wheat fields using an rgb-d camera. Information Processing in Agriculture **10**(4), 475–490 (2023). https://doi.org/10.1016/j.inpa.2022.05.002
23. Suri, S., Walmer, M., Gupta, K., Shrivastava, A.: Lift: A surprisingly simple lightweight feature transform for dense vit descriptors. In: European Conference on Computer Vision. pp. 110–128. Springer (2024)
24. Thota, M., Leontidis, G.: Contrastive domain adaptation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2209–2218 (2021). https://doi.org/10.1109/CVPRW53098.2021.00250
25. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Towards segmenting everything in context. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1130–1140 (2023)
26. Wang, Z., Li, F., Jiang, D., Li, L., Zhong, Y., Dai, G., Wang, M.: Unlocking the potential of grounding dino in videos: Parameter-efficient adaptation for limited-data spatial-temporal localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 8822–8831 (2026)
27. Weyler, J., Läbe, T., Magistri, F., Behley, J., Stachniss, C.: Towards domain generalization in crop and weed segmentation for precision farming robots. IEEE Robotics and Automation Letters **8**(6), 3310–3317 (2023). https://doi.org/10.1109/LRA.2023.3262417

28. Xiang, S., Guo, W., Burridge, J., Liu, S., Lu, H., Fukatsu, T.: Sprout: A scalable diffusion foundation model for agricultural vision. arXiv preprint arXiv:2603.27519 (2026), https://arxiv.org/abs/2603.27519
29. Zhang, A., Gao, G., Jiao, J., Liu, C.H., Wei, Y.: Bridge the points: Graph-based few-shot segment anything semantically. In: Advances in Neural Information Processing Systems. pp. 33232–33261 (2024)
30. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. In: International Conference on Learning Representations (2024)