SEMANTICS FROM STRUCTURE: UNSUPERVISED REPRESENTATION LEARNING FROM MULTI-VIEW IMAGES VIA DISTILLATION WITH FEATURE FIELDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised representation learning typically relies on heavy data augmentation and massive datasets. Reducing these dependencies is necessary for applications in domains where large-scale data collection is impractical. To this end, we introduce Semantics from Structure, a novel paradigm for unsupervised representation learning from multi-view observations that leverages 3D geometry to learn robust visual representations from limited data. Central to our methodology is Distillation with Feature Fields, which performs bidirectional distillation between a radiance field and a randomly initialized image encoder. We demonstrate that our framework successfully learns multi-view consistent semantic representations from a small collection of images without any augmentation or momentum encoders. The learned representations achieve strong downstream performance on 2D and 3D multi-view segmentation across a variety of scenes with no additional fine-tuning. Our approach demonstrates that learning from structural priors yields robust representations with drastically reduced data requirements.

1 INTRODUCTION

Visual representation learning is a key challenge in computer vision (Wiskott & Sejnowski, 2002; Hinton et al., 2006), as it enhances the performance and efficiency of downstream tasks (Oquab et al., 2014). Although self-supervised methods have made tremendous progress by discovering structure in unlabeled data (Caron et al., 2021), the prevailing methods heavily rely on extensive image datasets, sophisticated architectures, and aggressive data augmentation to encourage invariance. While these techniques are highly effective, they often ignore the underlying physical reality that images are 2D projections of the 3D world. Consequently, representations learned solely from 2D can fail with multi-view consistency and view-dependent occlusions, especially in video data where camera motion naturally provides multi-view cues.

Visual representation learning approaches that follow the Joint-Embedding Predictive Architecture (JEPA) paradigm exploit the natural progression of frames throughout videos to learn robust representations. Although effective, these methods remain dependent on huge video datasets and intricate training frameworks. For data abundant tasks, these techniques prevail. However, they struggle to generalize to tasks with scarce data such as high-throughput phenotyping in agriculture (Yin et al., 2025). Agriculture poses a unique challenge as collecting data for phenotyping requires specific seasons and plant maintenance (Miftahushudur et al., 2025). Other domains, such as deep sea and deep space exploration, face similar scarcity issues since data collection requires extremely expensive and specialized equipment.

To address these limitations, we introduce Semantics from Structure (SfS). SfS employs multi-view 3D geometry as a natural constraint for unsupervised representation learning from multi-view images, Figure 1. By explicitly modeling the 3D scene, SfS naturally enforces multi-view consistency without needing artificial augmentations. Core to this framework is Distillation with Feature Fields (DiFF). DiFF optimizes a feature field and a latent 3D Gaussian splatting (3DGS) scene representation, alongside a randomly initialized 2D image encoder. Through a process of bidirectional distillation, the 3D scene distills stable, view-consistent semantic features that bootstrap the 2D encoder, which in turn provides high-dimensional descriptors to the feature field. Leveraging the correspon-

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

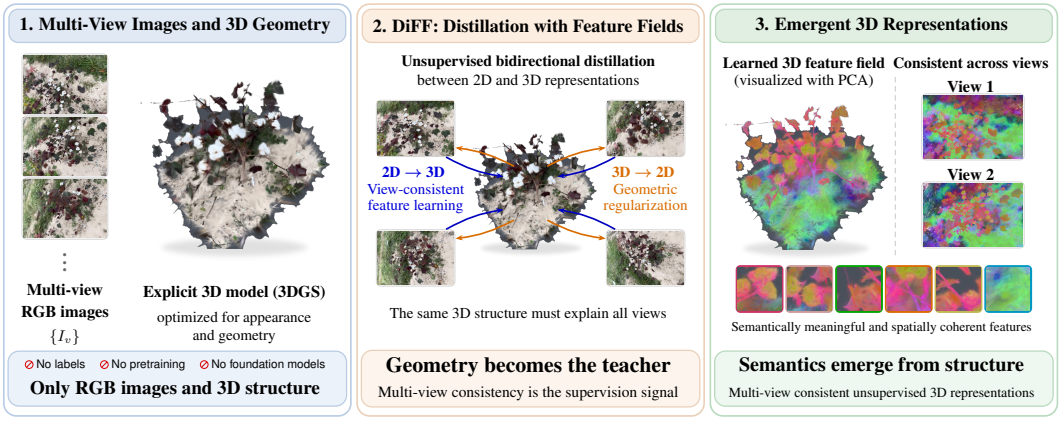


Figure 1: SfS provides emergent semantic representations from multi-view geometry as follows: 1. explicit 3D models optimized for photometric appearance and geometry; 2. unsupervised bidirectional distillation transfers multi-view geometric constraints between 3D and 2D representations; 3. discovered 3D latent feature fields, visualized via PCA, revealing multi-view consistent semantics directly from scene geometry without pretrained foundation models or labels.

dence between the static scene geometry and the true image observation, the learning objective is disambiguated, resulting in strong visual representations for downstream tasks with limited data.

We demonstrate that DiFF successfully learns robust, multi-view consistent semantic representations from a limited collection of images, completely avoiding data augmentation or momentum encoders typical of prior self-supervised methods. We evaluate DiFF on diverse downstream tasks including 2D and 3D multi-view segmentation across various scenes, demonstrating excellent performance and validating the efficacy of leveraging 3D structure for unsupervised learning. In summary, our contributions are as follows.

- We introduce a novel multi-view image representation learning paradigm that harnesses multi-view 3D geometry as a natural constraint.
- We propose a bidirectional distillation framework whereby the 3D scene distills view-consistent semantic features that bootstrap the 2D encoder, which in turn provides high-dimensional descriptors to the feature field.
- We design the spatio-semantic KNN smoothness loss to regularize and enforce a smoothness prior on the learned feature field by ensuring that the high-dimensional semantic features of points physically close in 3D space are highly similar.

The source code and multimedia associated with this project will be released after publication.

2 Related Work

Unsupervised Representation Learning. The goal of unsupervised visual representation learning is to extract meaningful semantic features without manual annotations. Joint-Embedding Architectures (JEAs) learn to produce similar embeddings for compatible inputs, typically generated via hand-crafted data augmentations (Chen et al., 2020). A primary challenge with JEAs is representation collapse, wherein the encoder learns to output a constant representation regardless of the input. Preceding work mitigated this by explicitly pushing apart negative examples using contrastive losses (Bromley et al., 1993; He et al., 2020; Chen & He, 2021), or by applying non-contrastive penalties that minimize informational redundancy (Bardes et al., 2021; Zbontar et al., 2021) and maximize entropy via clustering (Assran et al., 2021; 2022). Alternatively, methods such as BYOL (Grill et al., 2020) and DINO (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025) avoid collapse using asymmetric architectural designs, such as momentum encoders, a strategy later adapted for masked modeling by iBOT (Zhou et al., 2021) and extended to 3D distillation by Pri3D (Hou et al., 2021).

Conversely, generative approaches such as MAE (He et al., 2022) avert collapse by reconstructing raw pixels from masked inputs, which CroCo (Weinzaepfel et al., 2022) extends to multi-view image pairs via cross-view completion. JEPAs (LeCun, 2022; Assran et al., 2023) bridge these paradigms by applying the masked prediction task strictly in the embedding space. Since JEPAs seek predictive rather than purely invariant representations, they share the collapse risks of standard JEAs and rely on the same asymmetric encoder designs to prevent it. This predictive framework naturally extends to spatiotemporal data with models such as V-JEPA (Bardes et al., 2024), which learns by predicting the abstract representations of masked video tubes. By operating entirely in the embedding space, JEA and JEPA models efficiently capture complex appearance and motion features without the severe computational overhead of pixel-level decoding.

Novel View Synthesis and 3D Gaussian Splatting. Neural Radiance Fields (NeRFs) (Mildenhall et al., 2021) pioneered high-fidelity novel view synthesis by implicitly modeling scenes as continuous volumetric functions, mapping camera positions to appearance with multilayer perceptrons. 3DGS (Kerbl et al., 2023) has emerged as an efficient alternative, explicitly representing scenes with anisotropic 3D Gaussians and utilizing a fast tile-based rasterizer. Techniques based on 3DGS have surged in popularity due to its computational efficiency, fast training, and discrete geometric primitives. While 3DGS achieves real-time rendering and state-of-the-art visual quality, it primarily models appearance and geometry, lacking inherent semantic understanding.

Semantic Feature Fields. To bridge 3D geometry and semantics, several approaches distill powerful pretrained 2D foundational models into 3D representations. Distilled Feature Fields (Kobayashi et al., 2022), N3F (Tschernezki et al., 2022), and LERF (Kerr et al., 2023) distill CLIP (Radford et al., 2021) features into a semantic NeRF representation for open-vocabulary querying. LangSplat (Qin et al., 2024) and N2F2 (Bhalgat et al., 2024) extend this to 3DGS, achieving real-time open-vocabulary segmentation and language grounding. Frameworks such as Feature-3DGS (Zhou et al., 2024), SAGA (Cen et al., 2025), and SplatFill (Dahaghin et al., 2026) distill instance masks from SAM (Kirillov et al., 2023) into a 3DGS-based feature field. Complementarily, FiT3D (Yue et al., 2024) exploits 3DGS to enable 3D-aware adaptation of 2D foundation models.

FruitNeRF Meyer et al. (2024) and Wheat3DGS Zhang et al. (2025) demonstrate promising results towards applying feature fields to agricultural settings. Nonetheless, these methods are fundamentally supervised by foundation models that require large-scale pretraining. Early works (Florence et al., 2018) utilized 3D reconstructions for contrastive learning, while others have explored self-supervised object segmentation (Fan et al., 2022) and weakly supervised 3D scene segmentation (Liu et al., 2023; Castillo et al., 2024). Yet, these approaches still require pseudo-labels or simple scenes containing a single dominant object. Our work diverges significantly. We explore a fully unsupervised setting where both the 3D feature field and the 2D encoder are trained from scratch without any pretrained foundational models, relying strictly on the structure of the scene to learn robust semantic representations.

3 METHODOLOGY

DiFF is a cross-modality JEA that unifies 2D and 3D representations. It simultaneously learns a 3D semantic feature field and a 2D image encoder without any labels or pretrained weights. We build on 3DGS, which represents the scene with a set of 3D Gaussians characterized by means, scales, rotations, opacities, and colors. To enable semantic learning, we create a feature field using an adapted latent 3DGS (Zhou et al., 2025), where each Gaussian is augmented with a high-dimensional semantic feature vector. Figure 2 illustrates the overall architecture of DiFF.

3.1 LATENT 3D GAUSSIAN SPLATTING

In contrast to traditional latent 3DGS, which optimizes a latent 3D representation given a pretrained image encoder, our method learns a semantic latent space directly from the multi-view 3D geometry of the scene. The training pipeline consists of two sequential phases. First, camera extrinsics are obtained via COLMAP Schonberger & Frahm (2016) and a standard 3DGS is trained using photometric loss to capture the scene's geometry and appearance. Specifically, the scene is represented by a set of anisotropic 3D Gaussians $\mathcal{G}_i = \{\mu_i, \Sigma_i, \alpha_i, c_i, f_i\}$, where $\mu_i \in \mathbb{R}^3$ denotes the 3D position, $\Sigma_i \in \mathbb{S}^3_+$ is the 3D covariance matrix parameterized by scale $s_i \in \mathbb{R}^3$, and rotation unit quaternion

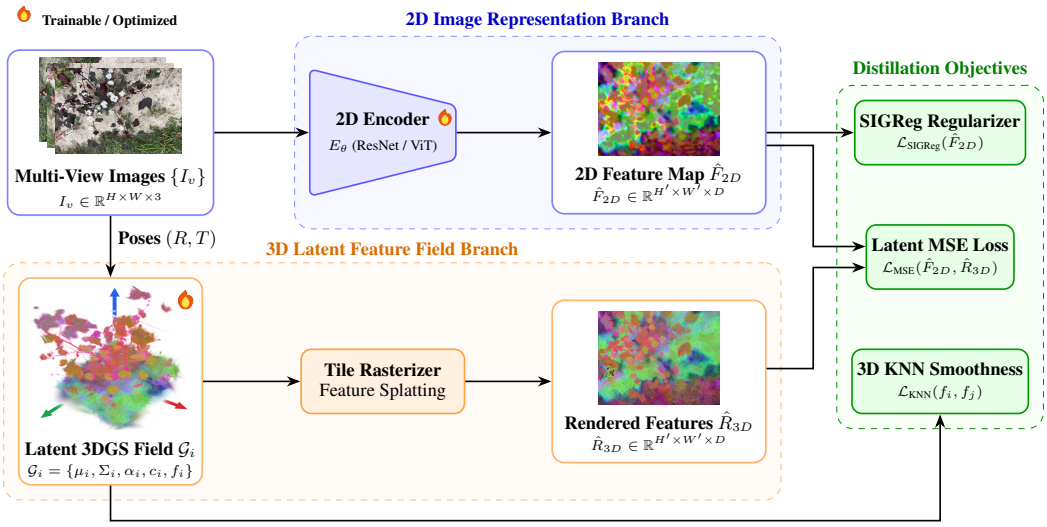


Figure 2: An overview of the DiFF framework. Input multi-view images are processed simultaneously by a 2D image encoder E_θ and a latent 3DGS field $\mathcal{G}_i$. Bidirectional feature distillation is guided by latent MSE matching ($\mathcal{L}_{\text{MSE}}$), SIGReg isotropic Gaussian regularization ($\mathcal{L}_{\text{SIGReg}}$) to prevent low-rank collapse, and 3D KNN spatio-semantic smoothness ($\mathcal{L}_{\text{KNN}}$).

$q_i \in \mathbb{H}$, $\alpha_i \in [0, 1]$ is the opacity, $c_i \in \mathbb{R}^3$ represents color appearance in RGB, and $f_i \in \mathbb{R}^D$ is the latent semantic feature vector of dimension D.

Once the geometry is stable, we freeze the geometric parameters $(\mu_i, \Sigma_i, \alpha_i)$ and colors c_i, and optimize the latent semantic features f_i alongside a randomly initialized 2D image encoder E_θ. During this latent training phase, we project and render the 3D Gaussian features to 2D via tile-based volume rendering,

$$R_{3D}(x) = \sum_{i\in\mathcal{N}} f_i \alpha'_i \prod_{j=1}^{i-1}(1-\alpha'_j), \tag{1}$$

where $\alpha'_i = \alpha_i \exp\left(-\frac{1}{2}(x-\mu_i^{2D})^T \Sigma_{i,2D}^{-1}(x-\mu_i^{2D})\right)$ represents the evaluated 2D opacity at pixel coordinate x, μ_i^{2D} and $\Sigma_{i,2D}$ are the projected 2D center and covariance matrix, and $\mathcal{N}$ is the depth-ordered set of Gaussians overlapping pixel x. Simultaneously, the 2D image encoder E_θ processes the normalized input image $I \in \mathbb{R}^{H\times W\times 3}$ to yield a 2D feature map $F_{2D} = E_\theta(I) \in \mathbb{R}^{H'\times W'\times D}$. To handle different encoder architectures, the backbone features are extracted, interpolated to a stride of 8, and mapped to latent dimension D using a single layer feed-forward network.

3.2 Loss Formulations for Latent Training

Latent training relies on three core loss functions to ensure the learned representations are consistent, well-distributed, and spatially coherent. The total latent objective is formulated as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{SIGReg}}\mathcal{L}_{\text{SIGReg}} + \lambda_{\text{KNN}}\mathcal{L}_{\text{KNN}}, \tag{2}$$

where loss balancing parameters λ_{SIGReg} and λ_{KNN} are specified in Section A.1.

Latent MSE Loss. We minimize the L2 distance between the normalized 2D encoder features and the normalized rendered 3D Gaussian features. The loss is formulated as

$$\mathcal{L}_{\text{MSE}} = \frac{1}{HW}\sum_i \|\hat{F}_{2D}(i) - \hat{R}_{3D}(i)\|_2^2, \tag{3}$$

where $\hat{F}_{2D}(i)$ and $\hat{R}_{3D}(i)$ denote the L2-normalized feature vectors at pixel location i.

Sketch Isotropic Gaussian Regularizer (SIGReg). To prevent feature collapse and enforce an isotropic Gaussian distribution across projected dimensions we adopt SIGReg (Balestriero & LeCun,

2025) regularization for the image encoder. Specifically, we evaluate the Epps-Pulley test statistic

$$\mathcal{L}_{\text{SIGReg}} = \frac{1}{P}\sum_{p=1}^{P}\int_t \left|\mathbb{E}[e^{it(X_p\cdot a_p)}] - \phi(t)\right|^2 w(t)dt, \tag{4}$$

where X_p represents a sampled subset of active pixel feature vectors, $a_p = A_{:,p}/\|A_{:,p}\|_2$ are unit random projections, $\phi(t) = \exp(-t^2/2)$ is the characteristic function of a standard normal distribution, and the integral is evaluated numerically over a 1D evaluation grid using a weighted integration window $w(t)$.

Spatio-Semantic KNN Smoothness Loss. We introduce a regularization term on the feature field to enforce a smoothness prior across the 3D space. The 3D feature field is regularized by explicitly encouraging spatial neighbors in 3D coordinate space to share similar semantic descriptors. Concretely,

$$\mathcal{L}_{\text{KNN}} = \frac{1}{|S|K}\sum_{i\in S}\sum_{j\in\mathcal{N}_K(i)} \|\hat{f}_i - \hat{f}_j\|_2^2, \tag{5}$$

where S is a sampled subset of Gaussian points, $\mathcal{N}_K(i)$ represents the K nearest neighbors of Gaussian i in 3D coordinate space under Euclidean distance $\|\mu_i - \mu_j\|_2$, and $\hat{f}_i$ is the L2-normalized Gaussian feature vector.

3.3 Representation Collapse Dynamics

The primary vulnerability in self-supervised visual representation learning is *representation collapse*, in which a model outputs constant descriptors to minimize distance-based objectives. Standard 2D self-supervised frameworks (Grill et al., 2020; Caron et al., 2021; Assran et al., 2022) require heuristic architectural mechanisms such as asymmetric predictor heads, stop-gradient operations, and momentum encoders to avoid collapse when optimizing distance functions such as MSE.

In our framework, the multi-view rendering constraint couples 2D features to a latent 3D Gaussian field, anchoring representations to physical coordinates $\mu_i \in \mathbb{R}^3$. While a constant field ($f_i = c, F_{2D}(x) = c$) is algebraically a trivial global minimum for normalized $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{KNN}}$, the 3D field ensures that any non-trivial features learned by the encoder must strictly satisfy 3D geometric and multi-view consistency across camera viewpoints, translations, and occlusions.

However, geometric rendering alone does not prevent representation collapse. In practice, optimizing solely under $\mathcal{L}_{\text{MSE}}$ leads the 2D encoder to suffer dimensional collapse, compressing representations onto a low-rank subspace with limited semantic discriminability. To rule out both trivial point collapse and low-rank dimensional collapse, we incorporate the statistical regularizer $\mathcal{L}_{\text{SIGReg}}$. By penalizing deviations from an isotropic Gaussian distribution, $\mathcal{L}_{\text{SIGReg}}$ enforces variance across all latent dimensions. This provides a principled statistical barrier against collapse, eliminating the need for architectural heuristics.

4 Downstream Tasks

The 2D and 3D representations learned by DiFF can be directly applied to various multi-view segmentation tasks without additional fine-tuning. We cluster the learned latent features to segment objects from novel views in 0-shot and 1-shot settings.

0-Shot 3D Unsupervised Clustering. We directly apply the normalized semantic features as a descriptor for each 3D Gaussian. The descriptors are clustered using k-means into $k = 8$ semantic categories based on Euclidean distance. The cluster assignments are mapped to one-hot vectors $C_i \in \{0,1\}^k$ and rendered via volume rasterization to generate 2D multi-class probability maps, followed by argmax label selection and morphological post-processing. The predicted masks are matched to ground truth via the Hungarian algorithm using intersection over union (IoU) as the assignment cost.

1-Shot 3D Prototype Rendering. Given a single annotated support view I_{supp} with foreground mask M_{supp}, we extract normalized 2D features F_{2D}^{supp} using the 2D encoder. The mask M_{supp} is

Table 1: Multi-view segmentation on the Wheat3DGS dataset. Subscripts indicate backbone initialization (IN = ImageNet pretrained). All methods use external data except where marked with †.

Method	Setting	IoU (%)	Prec. (%)	Rec. (%)	F1
FruitNeRF	Default	34.0	**95.0**	35.0	50.0
Wheat3DGS		50.0	81.0	57.0	67.0
DiFF†	0-Shot 3D	30.8	57.1	47.0	46.9
DiFF_{IN}		38.4	69.2	45.2	54.1
DiFF†	1-Shot 3D	40.2	57.8	58.2	56.8
DiFF_{IN}		46.9	68.2	59.5	63.2
LeJEPA†	1-Shot 2D	17.6	29.8	33.9	29.3
DiFF†		46.3	58.5	68.5	62.7
DiFF_{IN}		**53.5**	65.9	**72.2**	**68.6**

Table 2: Cross-scene 1-shot transfer evaluation on the Wheat3DGS dataset. The ResNet50 encoder is trained on `plot_461` and evaluated on unseen target plots. Retention (Ret.) represents the percentage of intra-scene 1-shot IoU preserved.

Target Scene	IoU (%)	Ret. (%)	Prec. (%)	Rec. (%)	F1
`plot_462`	43.2	94.5	58.2	62.8	60.4
`plot_463`	30.5	94.3	39.7	56.9	46.7
`plot_464`	56.4	98.5	70.4	73.9	72.1
`plot_465`	39.6	84.7	55.3	58.1	56.7
`plot_466`	52.4	100.7	62.8	76.0	68.7
`plot_467`	51.6	92.2	64.4	72.3	68.1
Mean	**45.6**	**94.3**	**58.4**	**66.7**	**62.1**

downsampled to a stride of 8 to match the resolution of the feature map, yielding positive and negative feature descriptors. We fit a balanced logistic regression classifier on the positive and negative samples to obtain classifier weight $W \in \mathbb{R}^D$ and bias $b \in \mathbb{R}$. For every 3D Gaussian i, we compute the decision logit $z_i = \hat{f}_i \cdot W + b$ and 3D probability score $p_i = \sigma(z_i) = \frac{1}{1+e^{-z_i}}$. We apply DBSCAN (Ester et al., 1996) to filter noise and remove spatial outliers. The 3D probability scores p_i are rendered to 2D test views and thresholded to obtain binary masks.

1-Shot 2D Feature Matching. We fit a classifier (W, b) on the support view identically to the 1-shot 3D method. For a novel test view I_{test}, the 2D feature map F_{2D}^{test} is generated directly by the 2D encoder. The decision boundary is applied in 2D, $z(x) = \hat{F}_{2D}^{\text{test}}(x) \cdot W + b$, and the score map $p(x) = \sigma(z(x))$ is upscaled to the resolution of the input image via bilinear interpolation and thresholded to yield semantic segmentation masks.

5 Experimental Results

To evaluate the effectiveness of DiFF, we performed comprehensive experiments across disparate scene types. First, we outline our evaluation protocol and the datasets (Section 5.1). Next, we report quantitative comparisons on the datasets (Section 5.2), evaluate multi-view feature consistency via cycle drift analysis (Section 5.4), and showcase qualitative segmentation overlays across challenging scenes (Section 5.5). Finally, we present extensive ablation studies (Section 5.6) to examine spatial regularization, training dynamics, network backbones, and resistance to representation collapse.

5.1 Datasets and Evaluation Benchmarks

To evaluate the generalization, multi-view consistency, and semantic segmentation capabilities of our unsupervised representations, we conducted experiments on two heterogeneous 3D multi-view benchmark datasets captured in distinct agricultural domains.

Wheat3DGS (Zhang et al., 2025) Dataset: Wheat3DGS is a field-based agricultural dataset captured in seven experimental wheat plot environments with each row related to a different wheat genotype. It poses severe challenges due to intricate canopy geometry, dense foliage occlusions, fine-grained plant structures, and natural lighting variations. The evaluation metrics used include IoU (%), precision (%), recall (%), and F1 score.

3DCotton (Al Muzaddid & Beksi, 2026a;b) Dataset: The 3DCotton dataset features high-resolution multi-view imagery and 3D reconstructions of cotton crops in the field. It provides SAM masks for cotton boll segmentation for all images in each scene. For the experiments, we utilized the plant_1 scene, consisting of 218 frames, to evaluate our unsupervised feature field distillation on intricate vegetative plant organs with fine-grained spatial distributions. We evaluated the quality of the cotton boll segmentation using IoU.

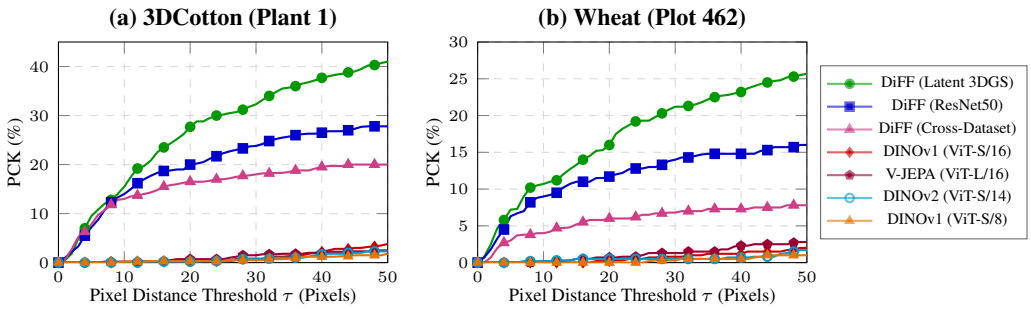


Figure 3: Multi-view cycle consistency percentage of correct keypoints (PCK) curves across (a) 3DCotton dataset (Plant 1) and (b) Wheat3DGS dataset (Plot 462) comparing DiFF against self-supervised 2D and video baselines (DINOv1, DINOv2, and V-JEPA). DiFF significantly reduces spatial cycle drift across diverse scenes, achieving up to 41.0% and 25.7% PCK (and 20.0% and 7.8% under cross-dataset transfer) at $\tau = 50$ px on the 3DCotton and Wheat3DGS datasets respectively, compared to SSL baselines (PCK $\leq 3.8\%$ and $\leq 2.8\%$).

5.2 Quantitative Results

We assessed DiFF on multi-view segmentation tasks to demonstrate the efficacy of its unsupervised representations. The results on the Wheat3DGS dataset are shown in Table 1. DiFF demonstrates strong capability in few-shot settings. The 1-shot 2D and 3D configurations outperform the supervised baseline, FruitNeRF, in terms of IoU and F1 score. The 1-shot 2D configuration approaches the performance of Wheat3DGS and even achieves a higher recall. Importantly, when training ResNet50 on the same scenes using LeJEPA (Balestriero & LeCun, 2025), performance degrades to 17.6% IoU due to severe data scarcity, demonstrating that 2D data augmentation alone cannot resolve viewpoint ambiguities. Even in the 0-shot unsupervised clustering setting, DiFF maintains strong performance, demonstrating that the distilled feature fields successfully capture inherent semantic boundaries within complex natural scenes. Furthermore, applying DiFF for transfer learning with pretrained ImageNet Deng et al. (2009) representations improves segmentation performance, achieving state-of-the-art results without requiring any masks for training.

5.3 Cross-Scene Generalization

The representations learned by DiFF can be directly applied towards downstream tasks on unseen scenes by utilizing the image encoder. To evaluate the transferability of the image encoder trained via DiFF, we conducted a cross-scene transfer experiment on the Wheat3DGS dataset. Specifically, we trained DiFF exclusively on a single scene, `plot_461` and the resulting encoder was evaluated on six unseen target plots under the 1-shot 2D setting. As shown in Table 2, the representations learned via DiFF demonstrate strong cross-scene generalization. Across all six unseen target plots, the encoder achieves an average IoU of 45.6%, retaining 94.3% of the performance attained by intra-scene models. Furthermore, the model maintains strong discriminative precision (58.4%) and recall (66.7%), yielding an overall mean F1 score of 62.1.

5.4 Multi-View Feature Consistency

To directly evaluate whether structural 3D distillation improves multi-view representation consistency, we measured cycle consistency keypoint drift across novel viewpoints on the 3DCotton and Wheat3DGS benchmarks. Specifically, given a query pixel in a source view, we establish nearest-neighbor feature correspondences across intermediate frames and compute the Euclidean pixel drift upon returning to the original view. Figure 3 plots the percentage of correct keypoints (PCK) as a function of the pixel distance error threshold $\tau \in [0, 50]$ pixels.

Standard 2D and video self-supervised models such as DINOv1, DINOv2, and V-JEPA suffer severe keypoint drift across camera poses (PCK $\leq 3.8\%$ on 3DCotton, $\leq 2.8\%$ on Wheat3DGS at $\tau = 50$ px). Notably, even though V-JEPA is trained on large-scale video data, its predictive objective in 2D embedding space fails to maintain spatial correspondences under camera motion and

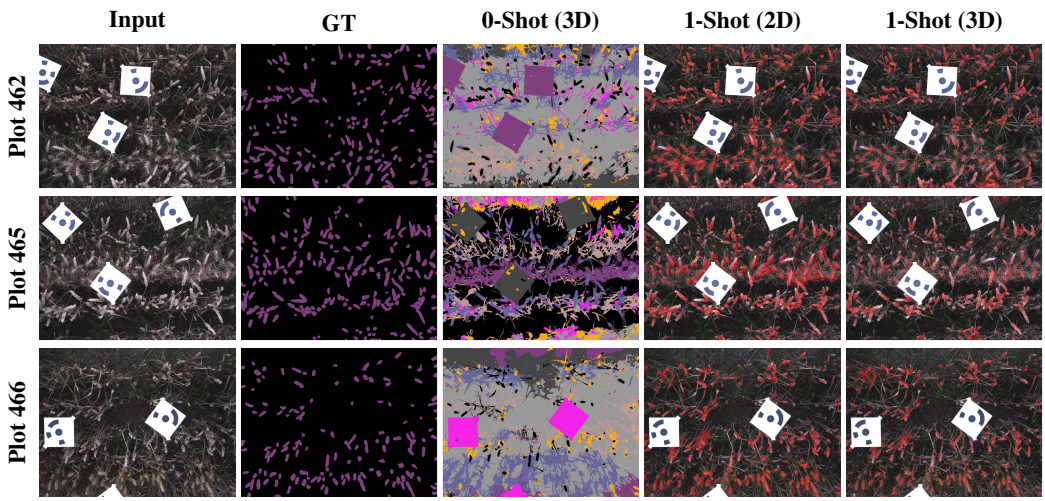


Figure 4: Qualitative semantic segmentation results on the Wheat3DGS dataset across 0-shot, 1-shot (2D), and 1-shot (3D) evaluation protocols. DiFF accurately delineates fine-grained plant structures and canopy boundaries under varying view angles without requiring manual annotations or pre-trained foundation models.

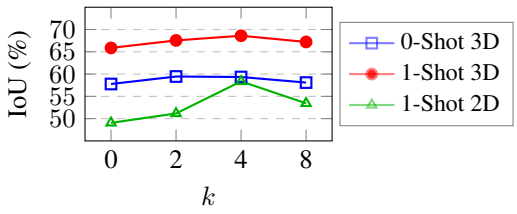


Figure 5: An ablation on k-nearest neighbors in $\mathcal{L}_{\text{KNN}}$ on the 3DCotton dataset. Enforcing spatial smoothness provides substantial improvements over unregularized baselines

Table 3: Best segmentation accuracy and corresponding k value for each method on the 3DCotton dataset.

Method	k	IoU (%)
0-Shot 3D	2	59.5
1-Shot 2D	4	58.4
1-Shot 3D	4	**68.6**

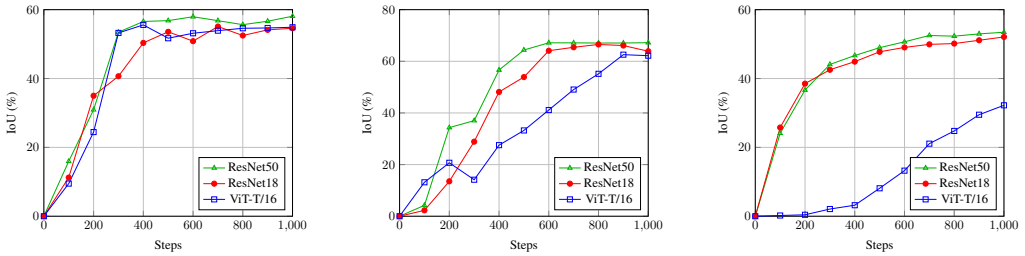
perspective shifts. On the other hand, representations distilled via DiFF achieve substantially higher correspondence accuracy, reaching a PCK of 41.0% on 3DCotton and 25.7% on Wheat3DGS for the 3D field (27.8% and 16.0% for the in-scene 2D encoder) at $\tau = 50$ px. In particular, when evaluated zero-shot across datasets, DiFF's encoder retains strong consistency (20.0% on 3DCotton, 7.8% on Wheat3DGS), outperforming all foundation baselines. This confirms that explicitly binding 2D features to an underlying 3D geometric field imparts strong multi-view consistency constraints that generalize across domains without artificial data augmentations.

5.5 Qualitative Results

Figure 4 presents qualitative segmentation results on representative scenes from the Wheat3DGS dataset across 0-shot 3D, 1-shot 2D, and 1-shot 3D evaluation settings. As shown, DiFF successfully produces fine-grained, view-consistent semantic segmentation masks that clearly isolate plant structures from background soil and clutter, even in the fully unsupervised 0-shot setting. Furthermore, the 0-shot visualizations show clear semantic understanding, as opposed to low-level appearance features, delineating wheat heads from visually similar stems.

5.6 Ablation Studies on 3DCotton

We performed extensive ablation experiments on the 3DCotton dataset to isolate the impact of our proposed core components, specifically evaluating 3D spatial smoothness regularization ($\mathcal{L}_{\text{KNN}}$), latent optimization step progression, and 2D image encoder backbones.



(a) 0-Shot 3D (b) 1-Shot 3D (c) 1-Shot 2D

Figure 6: Segmentation accuracy (mIoU vs. Latent Distillation Steps) across 2D image encoder architectures (ResNet50, ResNet18, ViT-T/16) on the 3DCotton dataset.

5.6.1 Effect of 3D k-Nearest Neighbors Smoothness

To quantify the impact of spatial regularization on the 3D feature field, we evaluated the segmentation accuracy (IoU) on 3DCotton across varying numbers of 3D nearest neighbors k used in $\mathcal{L}_{\text{KNN}}$. The optimal settings for k and the corresponding segmentation performance of each method are shown in Table 3. Figure 5 depicts the ablation trajectories across the 0-shot 3D, 1-shot 3D, and 1-shot 2D protocols.

Without 3D spatial regularization ($k = 0$), performance drops significantly across all evaluation modes on 3DCotton. As shown in Figure 5, setting $k = 4$ nearest neighbors achieves optimal performance across all evaluation modes, yielding substantial accuracy gains for 1-shot 2D and 1-shot 3D over unregularized baselines. Enforcing local feature coherence among neighboring Gaussians in 3D coordinate space prevents noisy, view-dependent feature inconsistencies. Nonetheless, larger neighbor radii ($k = 8$) cause mild feature over-smoothing across semantic object boundaries.

5.6.2 Training Dynamics and 2D Encoder Architectures

We analyzed the training dynamics of the latent distillation over optimization steps on 3DCotton for three 2D image encoder backbones: ResNet18 He et al. (2016), ResNet50, and ViT-T/16 (Dosovitskiy et al., 2020). Figure 6 compares the performance trajectories of each encoder for the 0-shot 3D, 1-shot 3D, and 1-shot 2D evaluation modes. All backbones demonstrate rapid learning during initial distillation, with ResNet50 stabilizing smoothly as optimization progresses. Convolutional backbones achieve higher final performance and faster convergence than ViT-T/16. In 1-shot 2D, ViT-T/16 requires more optimization steps to learn spatial correspondences without pretraining or spatial inductive biases, whereas ResNet50 achieves superior performance due to its inherent translation invariance and hierarchical spatial downsampling.

6 Conclusion

We introduced SfS via DiFF, a new paradigm for unsupervised representation learning from multi-view images that leverages 3D multi-view geometry as a natural constraint. By performing bidirectional distillation between 3D Gaussian feature fields and 2D image encoders, DiFF successfully learns view-consistent semantic descriptors from scratch without relying on pretrained foundation models, heavy 2D data augmentations, or momentum encoders. Our loss formulation combines latent MSE, SIGReg regularization, and 3D spatio-semantic KNN smoothness, ensuring semantically rich 2D and 3D feature representations. Downstream evaluations on 0-shot 3D unsupervised clustering and 1-shot novel-view segmentation on the Wheat3DGS and 3DCotton datasets confirm that structural multi-view constraints are highly effective for learning meaningful semantic visual representations. Future work includes expanding DiFF to multi-scene generalizable representations, incorporating temporal dynamic modeling for non-rigid video scenes, and application to multiple imaging modalities.

AI USE STATEMENT

In this work, generative AI tools were used to assist with source code development, generating TikZ code for figures, plots, and tables, and performing grammar, style, and technical manuscript editing. Generative AI was not used for narrative drafting, hypothesis generation, or scientific analysis. We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including content, claims, and code artifacts.

ETHICS STATEMENT

This work focuses on unsupervised representation learning from public multi-view image datasets. It does not involve human subjects, sensitive personal data, or harmful application domains.

REPRODUCIBILITY STATEMENT

To ensure complete reproducibility, all major hyperparameter settings, mathematical formulations, network architecture choices, and loss definitions are detailed in Section 3 and Section A.1.

REFERENCES

Md Ahmed Al Muzaddid and William J Beksi. Cropnerf: A neural radiance field-based framework for crop counting. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2026a.

Md Ahmed Al Muzaddid and William J Beksi. 3DCotton, 2026b. URL `https://doi.org/10.32855/dataset.2026.02.044`.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8423–8432, 2021.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *Proceedings of the European Conference on Computer Vision*, pp. 456–473. Springer, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Yash Bhalgat, Iro Laina, Joao F Henriques, Andrew Zisserman, and Andrea Vedaldi. N2f2: Hierarchical scene understanding with nested neural feature fields. In *Proceedings of the European Conference on Computer Vision*, pp. 197–214. Springer, 2024.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 6, 1993.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9630–9640, 2021.

Myrna Castillo, Mahtab Dahaghin, Matteo Toso, and Alessio Del Bue. Contrastive gaussian clustering for weakly supervised 3d scene segmentation. In *Proceedings of the International Conference on Pattern Recognition*, pp. 114–130. Springer, 2024.

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1971–1979, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, pp. 1597–1607. PmLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15745–15753, 2021.

Mahtab Dahaghin, Milind G Padalkar, Matteo Toso, Alessio Del Bue, and Vittorio Murino. Splatfill: 3d scene inpainting via depth-guided gaussian splatting. In *Proceedings of the International Conference on Pattern Recognition*, pp. 19–33. Springer, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, volume 96, pp. 226–231, 1996.

Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes. *arXiv preprint arXiv:2209.08776*, 2022.

Peter R Florence, Lucas Manuelli, and Russ Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. *arXiv preprint arXiv:1806.08756*, 2018.

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, Rémi Munos, and Michal Valko. Bootstrap your own latent-a new approach to self-supervised learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9726–9735, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15979–15988, 2022.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5673–5682, 2021.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139–1, 2023.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19672–19682, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3992–4003, 2023.

Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pp. 23311–23330, 2022.

Yann LeCun. A path towards autonomous machine intelligence version. *Open Review*, 62(1):1–62, 2022.

Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, pp. 53433–53456, 2023.

Lukas Meyer, Andreas Gilson, Ute Schmid, and Marc Stamminger. Fruitnerf: A unified neural radiance field based fruit counting framework. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1–8, 2024.

Tajul Miftahushudur, Halil Mertkan Sahin, Bruce Grieve, and Hujun Yin. A survey of methods for addressing imbalance data problems in agriculture applications. *Remote Sensing*, 17(3):454, 2025.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1717–1724, 2014.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20051–20060, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théeo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *Proceedings of the International Conference on 3D Vision*, pp. 443–453. IEEE, 2022.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pp. 3502–3516, 2022.

Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20310–20320, 2024.

Shuolei Yin, Yejing Xi, Xun Zhang, Chengnuo Sun, and Qirong Mao. Foundation models in agriculture: A comprehensive review. *Agriculture*, 15(8):847, 2025.

Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. In *Proceedings of the European Conference on Computer Vision*, pp. 57–74. Springer, 2024.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Daiwei Zhang, Joaquin Gajardo, Tomislav Medic, Isinsu Katircioglu, Mike Boss, Norbert Kirchgessner, Achim Walter, and Lukas Roth. Wheat3dgs: In-field 3d reconstruction, instance segmentation and phenotyping of wheat heads with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5360–5370, 2025.

Chaoyi Zhou, Xi Liu, Feng Luo, and Siyu Huang. Latent radiance fields with 3d-aware 2d representations. In *Proceedings of the International Conference on Learning Representations*, volume 2025, pp. 4933–4952, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21676–21685, 2024.

A Appendix

Additional architectural details, ablation studies, and qualitative segmentation overlays are provided in the sections below.

Table 4: Hyperparameter and architectural specifications for DiFF.

Hyperparameter / Setting	Value
Dataset and Views	
Training resolution	2048×1498 (Wheat); 1920×1440 (Cotton)
Frames per scene (train)	30 (Wheat); 191 (Cotton)
Views per step (batch size)	1
Distillation (Stage 2)	
Latent feature dimension D	16
2D backbone	ResNet-50
Optimizer & steps	Adam, 1,000 steps ($\text{lr}_{\text{3D}} = 10^{-4}$, $\text{lr}_{\text{2D}} = 10^{-3}$)
Rendering opacity threshold	$\alpha(p) \geq 0.1$
KNN neighborhood	$k = 4$, $\lvert S \rvert = 8{,}192$ points
KNN loss weight λ_{KNN}	1.0
SIGReg Regularizer	
1D random projections P	1,024 unit vectors on $\mathbb{S}^{D-1}$
Quadrature grid	17 knots on $[0, 3]$, $dt = 0.1875$
Quadrature window	Damped Gaussian $w(t) = \exp(-t^2/2)$
Subsampled active pixels M	$M \leq 1{,}024$
SIGReg loss weight λ_{SIGReg}	1.0×10^{-3}
1-Shot Classifier	
Logistic regression	L_2 penalty ($C = 1.0$), L-BFGS, max 200 iter
Class balancing	Balanced weights, negative sample $\leq 2 \times N_{\text{pos}}$
Decision thresholds τ	0.73 (2D), 0.715 (3D)
2D LeJEPA Baseline	
Training setup	1,000 steps, batch size 4
Backbone	ResNet-50
Multi-crop views V	$V = 4$ ($s \in [0.2, 1.0]$, jitter, blur, flip)
Projector	2-layer MLP ($16 \to 256 \to 128$), AdamW ($\text{lr} \in [10^{-5}, 10^{-3}]$)
Auxiliary loss weight λ	0.05 (dual projector + spatial SIGReg)
Cycle PCK Protocol	
Evaluation loop	3-hop closed loop ($I_0 \to I_8 \to I_{16} \to I_0$)
Query keypoints N	600 points on foreground mask
Search strategy	Global exhaustive search across feature map
Feature strides	1 (DiFF 3DGS), 8 (DiFF 2D), 8/14/16 (DINOv1/v2)

A.1 Implementation Details

Our framework is implemented in PyTorch and utilizes the `gsplat` CUDA library for tile-based rasterization. Training consists of two sequential phases, the details are as follows.

- **Stage 1: Geometry and Appearance Optimization.** We trained the 3DGS for 30,000 steps using the Adam optimizer to establish scene geometry and appearance following the standard 3DGS training recipe (Kerbl et al., 2023).
- **Stage 2: Latent Feature Distillation.** The geometric parameters (μ, Σ, α) and colors c are frozen. We set the latent dimension $D = 16$ and trained the latent features f_i and 2D encoder weights E_θ for 1,000 steps using the Adam optimizer with learning rates of 1.0×10^{-4} for the 3D latent features and 1.0×10^{-3} for the encoder components.

We set the tunable loss weights to $\lambda_{\text{SIGReg}} = 1.0 \times 10^{-3}$ and $\lambda_{\text{KNN}} = 1.0$. All experiments were conducted on a single consumer-grade GPU using a ResNet50 encoder with $k = 4$ nearest neighbors for 3D regularization, unless otherwise stated. Additional experiment details and parameters are provided in Table 4

A.2 Cross-Dataset Generalization

To further evaluate the generalization of the learned encoder representations, we performed a cross-dataset transfer experiment. Specifically, we trained DiFF on a single scene from 3DCotton and

Table 5: Cross-dataset 1-shot transfer evaluation from the 3DCotton to Wheat3DGS datasets. The ResNet50 encoder is trained on 3DCotton (`plant_1`) and evaluated zero-shot on unseen Wheat3DGS target plots. Retention (Ret.) represents the percentage of intra-scene 1-shot IoU preserved.

Target Scene	IoU (%)	Ret. (%)	Prec. (%)	Rec. (%)	F1
`plot_461`	32.1	95.3	44.4	53.6	48.6
`plot_462`	41.6	90.8	52.8	66.2	58.7
`plot_463`	31.1	96.2	37.1	65.7	47.5
`plot_464`	53.2	92.8	62.8	77.5	69.4
`plot_465`	37.5	80.2	50.6	59.1	54.5
`plot_466`	47.7	91.7	54.9	78.4	64.6
`plot_467`	47.2	84.3	57.1	73.1	64.1
Mean	**41.5**	**89.7**	**51.4**	**67.6**	**58.2**

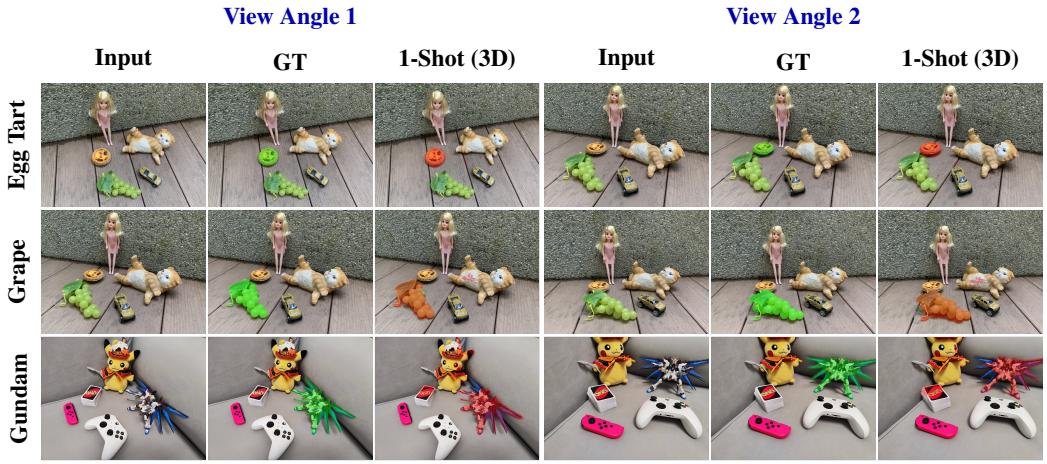


Figure 7: Qualitative 1-shot 3D segmentation results on the 3D-OVS dataset across target object categories (*Egg Tart*, *Grape*, *Gundam*) evaluated from two distinct camera view angles. DiFF demonstrates precise, multi-view consistent 3D object segmentation distilled without masks.

evaluated the multi-view segmentation performance on all Wheat3DGS plots using the 1-shot 2D setting. As shown in Table 5, the representations learned by the encoder are robust, enabling accurate segmentation performance in new environments. Across all plots, the encoder retains 89.7% of the IoU attained by intra-scene models, demonstrating strong transferability to new datasets.

A.3 Experiments on the 3D-OVS Dataset

We evaluated the 3D representations learned by DiFF on the 3D-OVS (Liu et al., 2023) dataset to validate the generalization ability of our bidirectional distillation framework on multi-class scenes. 3D-OVS provides five calibrated multi-view scenes of indoor and outdoor environments. Each scene contains between 20 and 40 images with high-quality ground-truth semantic mask annotations for multiple semantic classes for 5 selected frames.

In Table 6, we report quantitative segmentation results using 1-shot 3D prototype rendering on the 3D-OVS dataset. Notably, DiFF achieves accurate segmentation performance purely by discovering

Table 6: Multi-view semantic segmentation results on the 3D-OVS dataset (IoU %).

Method	Setting	bed	bench	room	sofa	lawn	mean
DiFF	1-shot 3D	61.3	64.8	53.5	63.1	58.7	60.3

Table 7: Quantitative verification of SIGReg. We report the effective rank (erank $\in [1, 16]$) of the representation covariance matrices along with downstream semantic segmentation accuracy (IoU %) on the 3DCotton dataset. Disabling $\mathcal{L}_{\text{SIGReg}}$ causes severe rank collapse (erank$_{2D} \approx 1.0$), rendering downstream segmentation non-functional.

Loss Components			**Covariance Metrics**			**Downstream IoU (%)**		
$\mathcal{L}_{\text{MSE}}$	$\mathcal{L}_{\text{KNN}}$	$\mathcal{L}_{\text{SIGReg}}$	erank$_{2D}$	λ_1 Var (%)	erank$_{3D}$	0-Shot 3D	1-Shot 3D	1-Shot 2D
✓	✓	✓	**15.0**	12.8	**15.9**	**59.5**	**68.6**	**58.4**
✓	×	✓	14.8	13.1	15.9	57.8	65.9	52.3
✓	✓	×	1.0	100.0	12.3	0.0	0.0	0.0
✓	×	×	1.0	100.0	12.5	0.0	0.0	0.0

structure from the existing multi-view geometry. Furthermore, Figure 7 illustrates the 1-shot 3D qualitative segmentation results on diverse indoor objects (*Egg Tart*, *Grape*, and *Gundam*) rendered from multiple camera viewpoints. Despite the complex semantic classes consisting of multiple parts, the 3D feature prototypes distilled by DiFF maintain precise object boundary selectivity and spatial consistency across changing view angles.

A.4 Representation Collapse and Objective Minimalism

To empirically validate our analysis in Section 3.3, we investigated model behavior on the 3DCotton dataset under minimal loss configurations (Table 7). When optimizing with an $\mathcal{L}_{\text{MSE}}$-only loss with both $\mathcal{L}_{\text{SIGReg}}$ and $\mathcal{L}_{\text{KNN}}$ disabled, the feature field never undergoes complete point collapse, maintaining an effective rank of 12.5. Instead, the multi-view rendering constraint forces the feature field to retain spatial variation across camera viewpoints. Nevertheless, under $\mathcal{L}_{\text{MSE}}$ alone, the encoder feature representations compress into a low-rank sub-manifold (erank$_{2D} = 1.0$, with λ_1 capturing 100.0% of the variance), reducing semantic discriminability. Incorporating SIGReg $\mathcal{L}_{\text{SIGReg}}$ on the 2D features restores dimensionality (erank$_{2D} = 14.8$, erank$_{3D} = 15.9$), while $\mathcal{L}_{\text{KNN}}$ enforces spatial smoothness in the 3D coordinate space, further improving downstream performance. This confirms that geometric constraints suffice to prevent 3D point collapse, while SIGReg prevents low-rank feature compression.

A.5 Qualitative 3D Latent Feature Field and Multi-View Visualizations

To further illustrate the representation quality and spatial consistency learned by DiFF, we visualize the 3D latent feature field on sophisticated plant organs present in the 3DCotton dataset across multiple viewpoints in Figure 8 (left column). The 3D latent feature field $f_i \in \mathbb{R}^{16}$ is projected into the RGB color space via 3-component principal component analysis (PCA). As shown in the middle column of Figure 8, the distilled feature field exhibits smooth intra-plant semantic feature continuity, effectively separating plant foliage and stems from the surrounding background terrain while preserving view consistency under novel camera perspectives. Furthermore, performing unsupervised 3D semantic clustering ($k = 8$) directly on the normalized 3D Gaussian descriptors (right column of Figure 8) produces sharp structural boundary separation between vegetative organs, soil, and ground clutter without requiring manual mask supervision or pretrained 2D foundation models.

A.6 Discussion and Future Work

Empirical results on downstream representations from DiFF demonstrate that robust semantics can be learned and extracted directly from the underlying 3D and 2D geometries present in a scene. Due to its simplicity, we believe that DiFF provides a strong prior to which future work can be built upon. The framework can be extended to multiple modalities, such as thermal, hyperspectral, and event-based vision, as the only fundamental requirement is a base 3D reconstruction and a 2D encoder method. Given the resulting clustered representations, the methodology could be further refined by incorporating a third training phase consisting of weakly supervised segmentation. Furthermore, the framework can be expanded for temporal representation learning by incorporating a 4D Gaussian splatting (Wu et al., 2024) method in place of the base 3DGS.

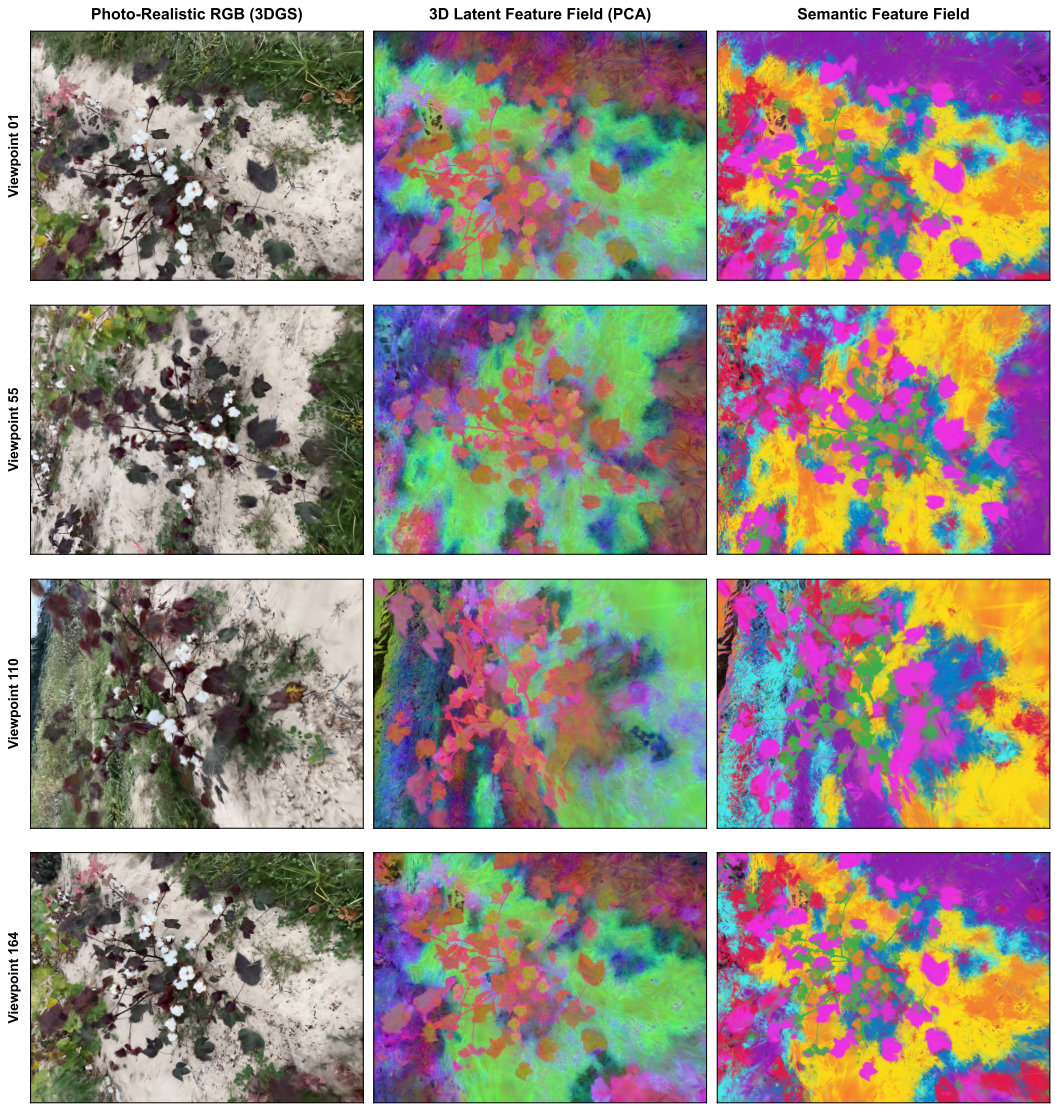


Figure 8: Qualitative reconstruction and 3D latent feature field representation for a cotton plant scene. Across multiple canonical camera viewpoints, we visualize the following: (left column) photo-realistic 3DGS RGB reconstruction; (middle column) 3D latent feature field projected via 3-component PCA into the RGB space, demonstrating smooth intra-plant and plant-ground semantic feature continuity; (right column) unsupervised 3D semantic clustering ($k = 8$), demonstrating clean structural boundary separation between foliage, stem, and background terrain.