DDMS: Discriminative Distillation of Multi-view Foundational Features into Single-view Models

Jeong-gi Kwak[1] **Sho Kagami**[2] **Yuki Ono**[3] **Kwang Moo Yi**[1]

[1]University of British Columbia
[2]Sony Semiconductor Solutions Corporation
[3]Sony Corporation

Abstract

Foundational visual features such as DINO have played a critical role across modern computer vision, and have recently become key components in multi-view feed-forward geometry estimators. In this work, we demonstrate that by re-distilling these multi-view models—their internal knowledge of 3D geometry—into a single-view estimator, we can obtain enhanced 3D consistent foundational features. Our key idea is to construct a *multi-view teacher* by fusing pretrained 2D foundation features with multi-view geometric features, and refining the fused representation with a discriminative ranking objective. Through our discriminative distillation framework, we enforce the learned features to be *both 3D consistent and locally distinctive*, while keeping them aligned with the feature space of the original foundation model to preserve the semantic structure of the pretrained representation. Consistency and local discriminability are critical for 3D computer vision problems such as forming semantic and geometric correspondences across images. To demonstrate the effectiveness of our method, we perform comprehensive experiments spanning multiple angles: direct feature analysis, dense prediction transfer, and explicit 3D lifting and rendering. Across these evaluations, our method consistently produces stronger 3D-aware foundation features that improve multi-view consistency and local discriminability while preserving the semantic transferability of the original representation.

1 Introduction

Recent vision foundation models [8, 44, 40, 53] pretrained on large-scale image data have emerged as powerful general-purpose visual backbones. Their features exhibit strong semantic transferability [1, 19, 12] and support a wide range of recognition and dense prediction tasks. However, because these features are primarily learned from single-image or image-level pretraining objectives [8, 20, 40], they often exhibit limited 3D awareness under multi-view evaluation [15]. In particular, features extracted from different views of the same scene or object may vary across viewpoints, while visually similar but geometrically distinct regions can remain insufficiently separated in feature space. Such inconsistency makes it difficult to reliably associate or aggregate features across viewpoints, limiting their effectiveness in 3D vision tasks that depend on stable cross-view feature associations.

To address this limitation, recent works [69, 68, 50] have sought to make pretrained visual features more consistent across views and more 3D-aware. A prominent direction uses 3D Gaussian Splatting (3DGS) [26] as an intermediate representation to construct multi-view feature supervision. Fit3D [69] lifts pretrained 2D foundation features into scene-specific 3D Gaussian representations and renders them from multiple views to supervise backbone fine-tuning. SnD [50] improves the scalability of this idea by replacing scene-specific optimization with a feed-forward 3DGS pipeline [9], distilling rendered 3D-aware features back into the backbone. These methods, however, encourage features

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

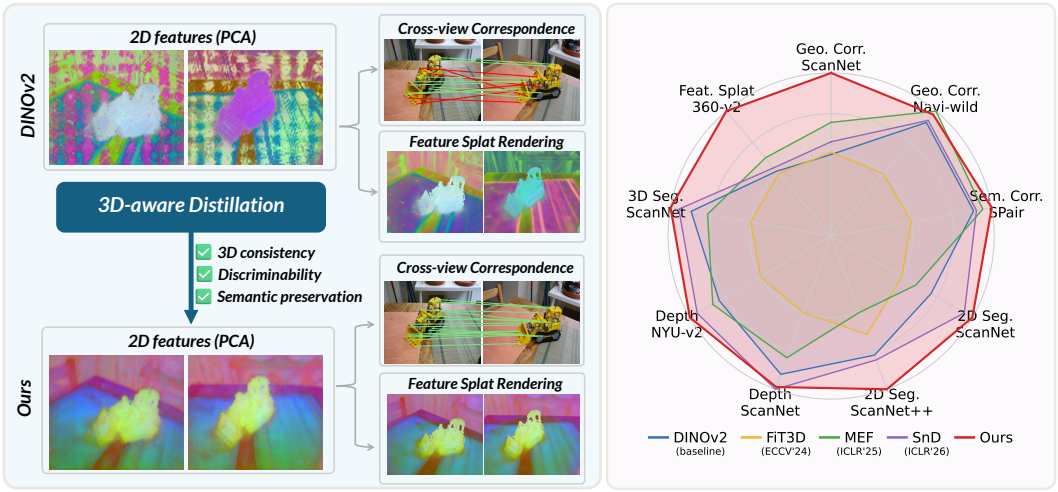

Figure 1: Our 3D-aware distillation turns pretrained 2D foundation features into representations that are more consistent across views while retaining their semantic utility. The resulting features produce cleaner cross-view correspondences and remain coherent after explicit 3D feature splatting (left). Across 3D-aware finetuning and distillation baselines [69, 68, 50] built from the same 2D foundation backbone [40], our representation provides a more balanced profile over geometric correspondence, semantic transfer, dense prediction, and 3D lifting evaluations (right). The radar chart axes use per-metric min–max normalization with the minimum set to 50%.

from different views to become similar, which can smooth out fine-grained distinctions that are critical for image matching, a foundational building block of 3D computer vision [33, 48, 47]. As shown in Figure 1, while the 3D consistency of semantic segmentation improves, geometric and semantic correspondence performances can be suboptimal, or even worse than the base foundational feature, as shown in the case of Fit3D [69].

Naturally, another line of work aims to improve feature discriminability, more in line with traditional correspondence-based descriptor learning [38, 34]. MEF [68] tunes pretrained features to serve as stronger matching descriptors for multi-view correspondence. Again, while this improves correspondence performance across views, it slightly weakens the semantic transferability of original foundational features as shown in Figure 1, making them more suitable as image correspondence features but not as general foundational features.

In this work, instead, we turn our attention to multi-view foundational models [61, 60, 30] that have recently shown impressive performance. Although these models are trained for geometric prediction rather than general-purpose feature extraction, their internal representations have been shown to encode correspondence and epipolar geometric cues [6], suggesting that they capture visual information useful for 3D reasoning. We therefore use these features together with foundational 2D features [8, 40] to derive a new representation that carries the best of both: the semantic utility of 2D foundation features and the geometric awareness of multi-view models. Then, to maximize utility at inference time, we distill the resulting *multi-view teacher* into a *single-view student*.

When learning the *multi-view teacher*, our goal is to jointly promote 3D consistency, fine-grained separability, and semantic preservation. To this end, we leverage the rich geometric priors of the multi-view model through a ranking-based objective, while anchoring the refined features to the original 2D foundation feature space. The ranking objective encourages features from the same local 3D surface to rank above geometrically incompatible candidates, rather than merely increasing similarity among all cross-view observations. At the same time, feature-space anchoring keeps the refinement aligned with the manifold of the original feature space (*e.g.*, DINO [8, 40]), helping preserve the semantic structure of the pretrained representation. After distillation, the resulting single-view features become more 3D-consistent, yield cleaner cross-view correspondences, and remain coherent after 3D feature lifting, as illustrated in Fig. 1. Through comprehensive experiments across multiple complementary evaluations in Sec. 4, we further show that these gains are not confined to a single aspect of the representation. Our method maintains strong performance across geometric, semantic,

and 3D aggregation settings, demonstrating a favorable balance between cross-view consistency, discriminability, and semantic preservation.

We summarize our main contributions as follows:

- we introduce a 3D-aware geometric foundational feature distillation framework that utilizes a multi-view teacher and a single-view student;
- to improve cross-view consistency while preserving separability and semantic structure, we propose to train the multi-view teacher via geometry-supervised ranking and feature-space anchoring; and
- with our single-view student, we demonstrate consistent gains across correspondence, dense prediction transfer, and explicit 3D lifting evaluations.

2 Related Work

Visual foundation models and 3D awareness. Recent visual foundation models [8, 40, 11, 71, 20, 44, 57, 45, 46, 53, 28] have shown strong transfer across recognition, dense prediction, segmentation, and correspondence-based tasks. A key reason for their broad utility is that their intermediate features often behave as dense visual representations rather than only image-level descriptors. This has enabled their use as dense visual descriptors [1], for unsupervised semantic grouping and segmentation [19, 12], as general-purpose dense features [40, 11, 17], and for point tracking or temporal correspondence [59]. Beyond 2D recognition and dense prediction, such features have also become important components in 3D vision pipelines, including dense image matching and correspondence estimation [13, 47, 54, 31, 14], visual odometry and SLAM [39, 56, 2], feature-based 3D lifting and feature field construction [58, 72, 36], and open-vocabulary or language-aligned 3D scene understanding [41, 27, 43]. These applications all rely, either explicitly or implicitly, on stable feature associations across views. However, generic visual foundation features are mostly learned from single images or image-level supervision and are not explicitly optimized for cross-view geometric consistency. Prior probing work [15] shows that such features contain useful but limited 3D awareness, with performance becoming less reliable under large viewpoint changes and multi-view correspondence settings. This limitation motivates recent efforts to adapt pretrained visual features with geometric, multi-view, or distillation-based supervision.

3D-aware feature learning and distillation. To address the limited 3D awareness of pretrained visual features, recent works have explored geometric supervision, 3D lifting, and distillation-based adaptation. Prior work has used multi-view or RGB-D geometry for representation learning [21, 35, 63, 64]; in contrast, DDMS adapts a pretrained 2D foundation model by constructing a teacher from frozen multi-view geometry features and distilling it into a single-view encoder. DUNE [46] studies a broader form of feature distillation by combining heterogeneous 2D and 3D teachers into a universal visual encoder. More directly related to our setting, several methods adapt a pretrained DINOv2 backbone [40] with explicit multi-view or 3D supervision. Fit3D [69] lifts 2D foundation features into scene-specific 3D Gaussian representations and renders them from novel views to fine-tune the backbone. SnD [50] improves scalability by replacing per-scene feature optimization with a feed-forward 3DGS pipeline, and incorporates SAM [28] masks for mask-aware feature upsampling and blending to strengthen dense semantic transfer. While effective for enforcing cross-view agreement, rendering- and aggregation-based supervision can smooth fine-grained feature differences, limiting local discriminability. MEF [68] takes a complementary direction by adapting pretrained features toward multi-view matching descriptors. This improves correspondence performance, but the resulting descriptor-oriented representation can become less suitable as a general-purpose dense feature.

Multi-view geometry and feed-forward 3D models. Recent multi-view geometry foundation models [61, 60, 30, 62, 25] have shifted 3D reconstruction from per-scene optimization toward generalizable feed-forward prediction. Given one or more images, these models predict geometric quantities such as depth, point maps, camera parameters, tracks, or metric scene structure using strong image encoders and cross-view reasoning. Feed-forward 3DGS methods [9, 65, 23] similarly produce renderable 3D representations without per-scene optimization. Beyond geometry prediction, 2D foundation features have also been lifted or aggregated into 3D representations for semantic and open-vocabulary scene understanding [55, 29, 66]. This feed-forward 3D paradigm has also been adopted for 3D-aware feature distillation. SnD [50], for example, renders feature supervision from a feed-forward 3DGS pipeline [9] to distill a 2D encoder. Rather than relying solely on rendered

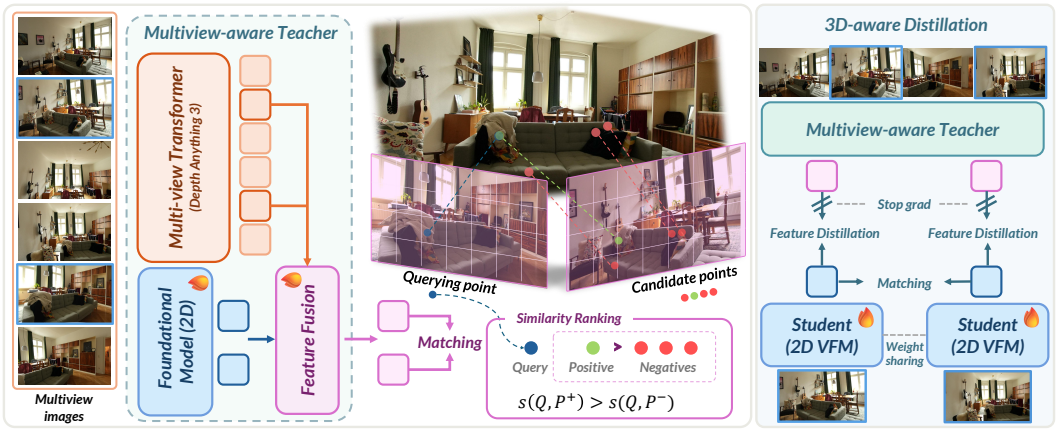

Figure 2: Overview of our method. We fuse pretrained 2D foundation features [40] with frozen multi-view geometry features [30] to construct a multi-view teacher. We refine the teacher with geometry-supervised ranking and feature-space anchoring, encouraging valid 3D correspondences to rank above incompatible candidates while preserving the original semantic feature space. We then distill it into a single-view student, which is used as the final feature extractor at inference time.

feature supervision, we use the intermediate features of a frozen multi-view geometry foundation model as multi-view geometric context for teacher construction. This choice is motivated by recent analyses showing that feed-forward 3D reconstruction models encode correspondence and epipolar cues in their intermediate representations [6]. This allows the teacher to combine geometric priors with the semantic structure of pretrained 2D features before distillation.

3 Method

Our goal is to adapt pretrained visual features into 3D-aware representations that improve cross-view consistency without sacrificing discriminability or the semantic richness of the original representation. To this end, we first construct a multi-view-aware teacher by enriching a 2D visual foundation model with geometric multi-view priors (Sec. 3.2), and further refine it through discriminative learning (Sec. 3.3). We then distill the resulting teacher into a student visual encoder (Sec. 3.4). The key idea is to inject multi-view geometric awareness into feature learning while maintaining discriminative matching ability and the semantic prior of the original 2D representation.

3.1 Overview

As illustrated in Fig. 2, our framework consists of two stages. First, we construct a multi-view-aware teacher by fusing features from a pretrained 2D visual foundation model with geometry-aware features from a multi-view transformer. The teacher is refined with a geometry-guided ranking objective, which ranks geometrically corresponding points above spatially distant 3D points in feature space. Second, we distill this teacher into a single-view student encoder using feature-level supervision. This allows the final student to be used like a standard 2D backbone at inference time, while inheriting multi-view geometric awareness from the teacher.

3.2 Multi-view-aware Teacher Construction

Given a set of N input views $\mathcal{I} = \{I_i\}_{i=1}^N$, we first run a geometric foundation model, implemented as a multi-view transformer, on all views jointly. Through cross-view attention, the model produces geometry-aware features whose representation for each view is contextualized by the full input set, allowing it to encode spatial relationships beyond what can be inferred from a single image alone. Feature learning is performed on a subset of views $V \subset \{1, \ldots, N\}$ with $|V| = M$. For each selected view $m \in V$, a pretrained 2D visual foundation model processes I_m independently and produces semantic features F_m^v, while we extract an intermediate feature F_m^g from the multi-view transformer [30], where correspondence and epipolar cues are known to emerge [6]. Since F_m^g is

optimized for geometric prediction rather than general-purpose visual representation, we leverage it as geometric context for refining the semantically rich 2D feature F_m^v. The two representations are concatenated on the same ViT patch grid. A lightweight projection and refinement module $\psi(\cdot)$ maps the concatenated feature $[F_m^v; F_m^g]$ back to the original foundational feature space and predicts a residual update:

$$T_m = F_m^v + \psi([F_m^v; F_m^g]), \qquad m \in V, \tag{1}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation. This residual feature fusion injects cross-view geometric context into the semantic representation while keeping the teacher close to the original feature space. We refine the teacher by optimizing the fusion module and LoRA adapters [22] on the 2D backbone, while keeping the geometry model frozen.

3.3 Discriminative Feature Learning

The fused teacher from Sec. 3.2 combines semantic 2D features with multi-view geometric context, but feature fusion alone does not guarantee local feature separability. We therefore encourage geometrically valid correspondences to rank above incompatible candidates in feature-similarity space. For each query pixel $x \in \mathcal{Q}$, we sample a candidate set $\mathcal{C}(x)$ from other views using the available multi-view geometry. Unreliable candidates are removed using reprojection validity and visibility/depth consistency. Among the remaining candidates, binary relevance labels are assigned by 3D proximity: candidates with $\|X_x - X_y\|_2 < \tau_{\text{pos}}$ are positives, candidates with $\|X_x - X_y\|_2 > \tau_{\text{neg}}$ are negatives, and candidates in between are treated as ambiguous and removed from $\mathcal{C}(x)$. Let $r_x(y) \in \{0, 1\}$ denote the relevance label of the resulting candidate set and $s(x, y) = \cos(t_x, t_y)$ denote the teacher-feature similarity. We optimize this ranking constraint using a differentiable average-precision objective [7, 68]. We approximate the rank of each candidate $y \in \mathcal{C}(x)$ as

$$R_x(y) = 1 + \sum_{\substack{z \in \mathcal{C}(x) \\ z \neq y}} \sigma\left(\frac{s(x, z) - s(x, y)}{\tau}\right), \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function and τ controls the softness of the rank approximation. The positive rank is computed using only relevant candidates:

$$R_x^+(y) = 1 + \sum_{\substack{z \in \mathcal{C}(x) \\ z \neq y}} r_x(z)\sigma\left(\frac{s(x, z) - s(x, y)}{\tau}\right). \tag{3}$$

We compute the ranking loss only over query pixels with at least one valid positive candidate, $\mathcal{Q}^+ = \{x \in \mathcal{Q} \mid \sum_{y \in \mathcal{C}(x)} r_x(y) > 0\}$. The discriminative loss is

$$\mathcal{L}_{\text{disc}} = 1 - \frac{1}{|\mathcal{Q}^+|} \sum_{x \in \mathcal{Q}^+} \frac{1}{Z_x} \sum_{y \in \mathcal{C}(x)} r_x(y) \frac{R_x^+(y)}{R_x(y)}, \qquad Z_x = \sum_{y \in \mathcal{C}(x)} r_x(y) + \epsilon. \tag{4}$$

This loss penalizes cases where geometrically incompatible candidates are ranked above valid correspondences. To allow geometric refinement while bounding deviation from the original 2D feature space, we use a semantic anchor to the frozen foundational feature v_x:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{|\mathcal{Q}|} \sum_{x \in \mathcal{Q}} \max\left(0, d_{\cos}(t_x, v_x) - \delta\right), \tag{5}$$

where $d_{\cos}(a, b) = 1 - \cos(a, b)$ and δ allows limited deviation from the original representation. The teacher refinement objective is

$$\mathcal{L}_{\text{teacher}} = \mathcal{L}_{\text{disc}} + \lambda_{\text{anchor}} \mathcal{L}_{\text{anchor}}. \tag{6}$$

3.4 Distilling Multi-view Priors into a Single-view Student

The refined teacher relies on multi-view inputs and geometry-based positive and negative candidates during training, whereas our final goal is a practical encoder that operates on a single image. After teacher refinement, we freeze the teacher and distill its representation into a single-view student. For each selected view $m \in V$, the student independently processes the image and produces a feature map

	Method	Separation	Geometric correspondence				Semantic matching		
		ScanNet [10]	ScanNet [10]		Navi-W [24]		DAVIS [42]	SPair-71k [37]	
		Margin↑	@10px↑	@20px↑	@0.05↑	@0.1↑	J&F↑	@0.05↑	@0.1↑
ViT-B	DINOv2 (baseline)	0.4108	21.72	36.11	57.97	75.94	67.82	52.98	68.56
	Fit3D (ECCV'24)	0.3494	22.89	39.50	31.26	53.81	66.70	30.25	44.86
	MEF (ICLR'25)	0.4288	39.17	55.06	**67.71**	**80.87**	66.82	56.78	72.07
	SnD (ICLR'26)	0.4520	28.65	45.67	59.53	76.86	68.05	53.68	69.65
	SnD (+blending)	0.3526	24.37	40.79	56.33	75.11	**69.11**	52.25	68.14
	Ours	**0.4632**	**66.01**	**78.48**	65.66	79.40	68.95	**61.52**	**75.46**
ViT-S	DINOv2 (baseline)	0.3967	23.00	38.16	53.72	73.03	65.78	51.11	67.16
	Fit3D (ECCV'24)	0.3001	22.27	39.29	26.60	47.87	66.01	21.12	34.21
	MEF (ICLR'25)	0.4170	38.54	54.30	**61.67**	**76.31**	66.16	51.64	67.32
	SnD (ICLR'26)	0.3819	28.65	46.24	54.13	73.30	67.10	51.13	67.83
	SnD (+blending)	0.3147	23.26	39.39	50.22	70.42	**68.20**	49.43	66.17
	Ours	**0.4314**	**61.35**	**74.24**	59.57	74.55	67.68	**54.09**	**68.70**

Table 1: Direct feature evaluation compared with recent 3D-aware feature finetuning methods [69, 68, 50]. Benchmarks cover ScanNet [10] feature separation and geometric correspondence, Navi-Wild [24] object correspondence, SPair-71k [37] semantic keypoint matching, and DAVIS [42] mask propagation. Best and second-best results are highlighted within each backbone group. Our method achieves the best separation, outperforms existing methods by a large margin on ScanNet, and performs strongly in all cases.

$S_m = f_\theta(I_m)$ in the same feature space as the teacher. We supervise the student with feature-level distillation from the refined teacher while stopping gradients through the teacher:

$$\mathcal{L}_{\text{distill}} = \frac{1}{|V|} \sum_{m \in V} \frac{1}{|\Omega_m|} \sum_{x \in \Omega_m} \frac{1}{D} \left\| \bar{S}_m(x) - \text{sg}\big[\bar{T}_m(x)\big] \right\|_1, \tag{7}$$

where $\bar{S}_m(x)$ and $\bar{T}_m(x)$ are the L2-normalized student and teacher features, D is the feature dimension, Ω_m denotes the valid pixel set, and $\text{sg}[\cdot]$ denotes stop-gradient.

To preserve the discriminative structure learned by the teacher, we also apply the same geometry-supervised ranking objective from Sec. 3.3 to the student features as a weak regularizer. The full student objective is

$$\mathcal{L}_{\text{student}} = \mathcal{L}_{\text{distill}} + \lambda_{\text{disc}} \mathcal{L}_{\text{disc}}, \tag{8}$$

where $\mathcal{L}_{\text{disc}}$ is computed using the student features S_m, and λ_{disc} controls the strength of the discriminative regularization. At inference time, only the student encoder is used, enabling standard single-image feature extraction with multi-view awareness distilled from the teacher.

4 Experiments

We evaluate DDMS from three complementary perspectives: direct feature quality, dense prediction transfer, and explicit 3D lifting and aggregation. Together, these experiments test whether our representation improves multi-view consistency and discriminability while preserving the semantic utility of pretrained visual features. We compare against recent 3D-aware feature learning baselines and ablate the main components of our method.

4.1 Experimental Setup

Benchmarks and evaluation. We first evaluate feature quality directly through feature separation, geometric and semantic correspondence, and temporal propagation on ScanNet [10], Navi-Wild [24], SPair-71k [37], and DAVIS-2017 [42]. Second, we evaluate downstream transferability with frozen-feature linear probes for semantic segmentation on ScanNet and ScanNet++ [67], and depth estimation on ScanNet and NYUv2 [52]. Third, we evaluate the features after explicit 3D lifting in two complementary settings. To isolate cross-view agreement, we perform point-based feature aggregation for ScanNet 3D point segmentation. We then lift image features onto pretrained 3D Gaussians [26, 36] and render them from held-out views on the Mip-NeRF 360 scenes [4]. Detailed evaluation protocols are provided in the supplementary appendix.

Baselines. For the main comparisons, we use DINOv2 [40] as the common pretrained 2D backbone, since closely related 3D-aware feature finetuning and distillation methods provide DINOv2-based

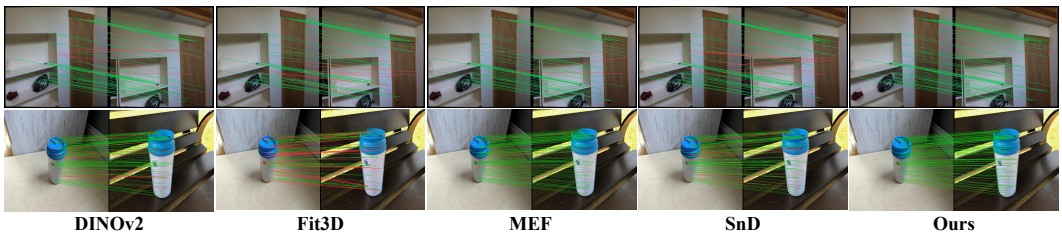

Figure 3: Qualitative nearest-neighbor correspondence results on ScanNet [10] and Navi-Wild [24]. Green and red lines denote correct and incorrect matches, respectively. Our features produce cleaner correspondences with fewer mismatches in these examples.

checkpoints. We compare against frozen DINOv2, Fit3D [69], MEF [68], and SnD [50]. For SnD, we report both released checkpoints: the checkpoint without SAM-mask blending, denoted SnD, and the checkpoint with blending, denoted SnD (+blending). This controlled setting keeps the initial feature space fixed and isolates how each method adapts the same pretrained representation for 3D-aware feature learning.

4.2 Geometric and Semantic Feature Quality – Table 1 and Figs. 3 and 4

We first evaluate the learned representation in feature space, focusing on separation, geometric correspondence, semantic correspondence, and temporal propagation.

Feature separation. To measure feature distinctiveness, we compute the margin between the cosine similarity of each corresponding pair and the average similarity to non-corresponding points. As shown in Table 1, our method achieves the largest margin across both backbone variants, indicating enhanced separation.

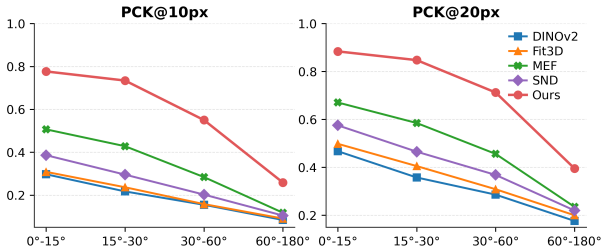

Figure 4: Viewpoint-gap analysis for ScanNet [10] correspondence (ViT-B, PCK@10px and PCK@20px). Our method remains robust as viewpoint changes increase.

Geometric correspondence. We first evaluate multi-view matching on ScanNet [10], where image pairs of indoor scenes are provided, and we measure the accuracy of the nearest-neighbor matches in feature space—the accuracy reflects geometric alignment. As shown in Table 1, our method achieves much higher correspondence accuracy on ScanNet. Qualitative results in Figure 3 also show more accurate matches with fewer errors. The viewpoint-gap analysis in Fig. 4 further shows that our advantage persists as the viewpoint gap increases.

Semantic correspondence. We then test whether this improvement transfers beyond shared-scene geometry in Table 1. Navi-Wild [24] evaluates object-centric correspondence under appearance and context changes, while SPair-71k [37] measures semantic correspondence across object instances. We also report DAVIS-2017 [42] as an auxiliary temporal propagation benchmark. Across these settings, our method remains consistently strong, suggesting that the gains in geometric correspondence and feature separation do not come at the expense of semantic correspondence.

4.3 Dense Prediction Transfer via Linear Probing – Table 2 and Figs. 5 and 6

We next evaluate whether the learned features preserve downstream utility beyond correspondence-based tasks by performing linear probing for semantic segmentation and depth estimation.

Semantic segmentation. To probe semantic richness, we follow the standard protocol of fitting a linear transformation from our features to semantics. As shown in Fig. 5, our method produces clean segmentation maps, consistent with the quantitative results in Table 2. Our method performs

	Method	Semantic segmentation						Depth estimation					
		ScanNet [10]			ScanNet++ [67]			ScanNet [10]			NYUv2 [52]		
		mIoU↑	mAcc↑	aAcc↑	mIoU↑	mAcc↑	aAcc↑	AbsRel↓	RMSE↓	δ_1 ↑	AbsRel↓	RMSE↓	δ_1 ↑
ViT-B	DINOv2 [40]	53.69	66.29	77.63	31.69	41.41	81.47	0.1553	0.3718	78.12	0.1705	0.6566	74.90
	Fit3D [69]	51.35	63.37	76.96	30.58	38.31	84.06	0.2045	0.4599	66.70	0.1975	0.7426	68.03
	MEF [68]	52.49	65.17	76.93	29.40	38.62	81.43	0.1677	0.3937	75.06	0.1661	0.6222	75.98
	SnD [50]	56.36	69.40	79.43	31.95	41.67	82.62	**0.1449**	**0.3446**	**80.89**	0.1537	0.6018	78.56
	SnD [50] (+blending)	**57.64**	70.35	80.52	33.37	42.73	84.13	0.1454	0.3459	80.54	**0.1464**	**0.5777**	**80.70**
	Ours	57.00	**70.51**	**80.61**	**33.53**	**43.00**	**84.34**	0.1450	0.3453	80.48	0.1490	0.5835	79.78
ViT-S	DINOv2 [40]	50.18	62.81	75.06	27.80	37.06	79.52	0.1709	0.3950	74.70	0.1699	0.6465	75.26
	Fit3D [69]	44.62	56.41	72.56	26.19	33.49	81.97	0.2171	0.4821	64.44	0.2120	0.7691	65.20
	MEF [68]	47.22	59.70	73.29	24.53	32.96	78.69	0.1925	0.4319	69.72	0.1792	0.6541	73.39
	SnD [50]	51.38	64.11	76.10	27.90	36.90	81.00	**0.1612**	0.3780	**76.45**	0.1536	0.5882	79.32
	SnD [50] (+blending)	**52.61**	**65.23**	77.07	29.18	38.14	**82.40**	0.1631	**0.3777**	76.22	**0.1531**	**0.5842**	**79.46**
	Ours	52.23	64.96	**77.12**	**29.29**	**38.53**	82.17	0.1628	0.3786	76.25	0.1582	0.6036	78.13

Table 2: Linear probing comparisons. We evaluate semantic segmentation on ScanNet [10] and ScanNet++ [67], and depth estimation on ScanNet [10] and NYUv2 [52]. Our method is best or second-best on every segmentation metric and remains competitive with the SnD variants [50] on depth, while outperforming both variants on ScanNet geometric correspondence in Table 1.

comparably to SnD (+blending) in semantic segmentation, while outperforming both SnD variants on ScanNet geometric correspondence in Table 1. DDMS therefore retains strong dense semantic transfer while improving geometric correspondence.

Depth estimation. To evaluate whether the learned representation retains single-view geometric cues, we train a linear depth probe on the frozen features. As shown in Fig. 6, our method produces coherent depth structure, consistent with the quantitative results in Table 2. DDMS remains competitive with SnD (+blending) on depth estimation while outperforming both SnD variants on ScanNet correspondence in Table 1. These results indicate that the learned representation retains useful geometric cues while improving cross-view discriminability.

4.4 Evaluating Features through 3D Lifting and Rendering – Table 3 and Figs. 7 and 8

We further test whether the learned features remain consistent and discriminative after explicit 3D lifting and rendering.

3D point segmentation. To directly look into 3D agreement, we first evaluate semantic transfer after point-based 3D aggregation on ScanNet [10]. We project image features directly onto 3D point clouds using posed RGB-D observations, average them across valid views, and train a linear classifier on the frozen aggregated features. Since averaging over many views can hide disagreement between them, we also apply the same classifier to the feature of each individual view and measure how often two views of the same point agree. As shown in Tab. 3, our method matches the strongest baseline after aggregation and performs best per view, while yielding the highest agreement across views. Importantly, this improved multi-view agreement is achieved while retaining the strong correspondence and dense-transfer performance observed in Tabs. 1 and 2.

3D feature splatting. We next evaluate the quality of features under a neural rendering setup. Specifically, we evaluate feature splatting on pretrained 3D Gaussian scenes [26, 4]. Using a fixed

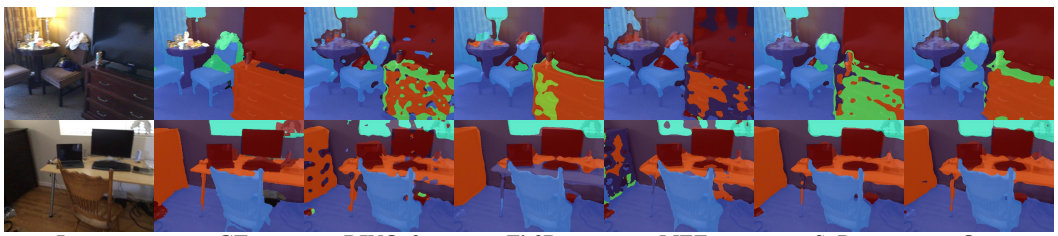

Figure 5: Qualitative comparison of semantic segmentation using frozen features with linear probing.

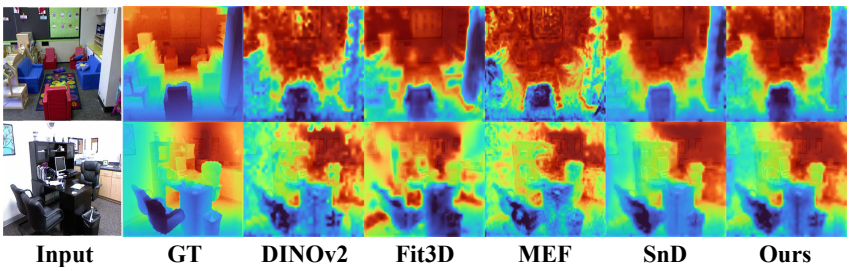

Figure 6: Qualitative comparison of depth estimation using frozen features with linear probing. Our method provides coherent depth estimates.

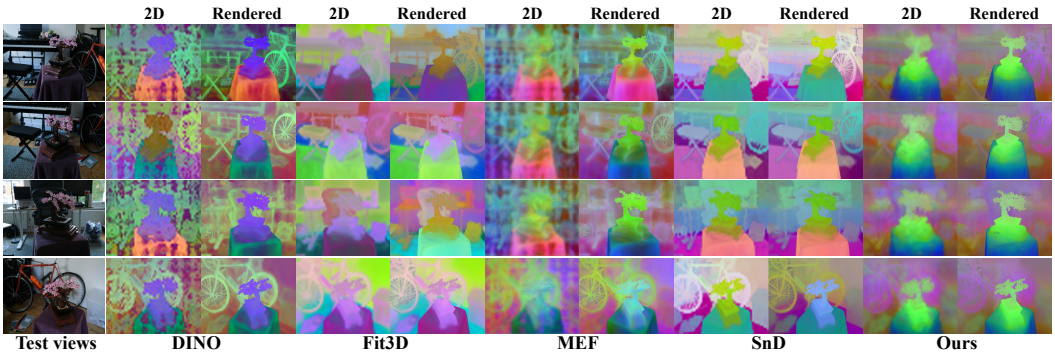

Figure 7: Qualitative examples of feature uplifting and rendering on pretrained 3D Gaussians [4]. PCA-projected 2D features are compared with rendered features using the same uplifting method [36]. Our features remain more coherent across views and after rendering.

uplifting method [36], we assign image features to the already-trained 3D Gaussians of each scene and render the resulting feature field from held-out views. As shown in Fig. 7, several baselines show discrepancies between their original 2D features and rendered features after splatting, suggesting cross-view mismatch or smoothing during Gaussian feature accumulation. In contrast, our rendered features remain better aligned with the corresponding 2D features and preserve coherent local structures. We quantify this by measuring the accuracy of nearest-neighbor matches when matching between image features (2D) and rendered features in Fig. 8. As expected, matching across rendered features shows high consistency for all methods, while these drop when matching rendered features to image features. Ours retains the highest accuracy in all cases, demonstrating consistency between image features and uplifted features.

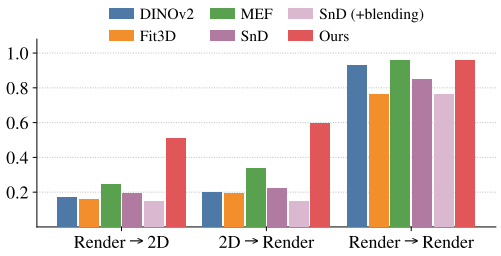

Figure 8: Feature consistency between image features and rendered features, measured as nearest-neighbor matching accuracy (PCK@10px) when matching across them. Our method retains the highest accuracy, demonstrating consistency.

4.5 Ablation study – Table 4

To motivate our design choices, we report ablations in Tab. 4, using ScanNet [10] feature separation and geometric correspondence, SPair-71k [37] semantic correspondence, and ScanNet semantic segmentation transfer. Directly fine-tuning DINOv2 with $\mathcal{L}_{\text{disc}}$ yields the strongest feature separation and geometric correspondence, but substantially degrades semantic correspondence and segmentation. This reveals a clear trade-off: optimizing discriminability alone can over-specialize the representation toward geometric matching at the expense of its general semantic utility. Adding $\mathcal{L}_{\text{anchor}}$ to direct fine-tuning recovers much of the lost semantic performance while retaining strong geometric correspondence, but remains below the full DDMS model across all four metrics. Starting from the full DDMS model, removing the teacher $\mathcal{L}_{\text{anchor}}$, the student $\mathcal{L}_{\text{disc}}$, or multi-view context

Method	mIoU↑	mAcc↑	aAcc↑	Per-view mIoU↑	Multi-view agree.↑
DINOv2 [40]	62.30	74.30	82.69	54.49	77.34
Fit3D [69]	59.54	71.04	81.45	54.21	80.73
MEF [68]	61.54	73.73	82.24	54.07	77.71
SnD [50]	62.79	74.57	82.99	56.51	80.28
SnD [50] (+blending)	63.26	**75.18**	83.25	57.14	80.80
Ours	**63.32**	74.85	**83.36**	**57.79**	**82.49**

Table 3: 3D point segmentation on ScanNet [10] with a linear probe on features aggregated over views. Per-view mIoU applies the same probe to the feature of each view without averaging, and multi-view agreement is the rate at which two views of the same point predict the same class. Ours matches the strongest baseline after aggregation and is best per view, with the most consistent predictions across views.

Variant	Sep. ↑	Corr. ↑	Sem. ↑	Seg. ↑
DINOv2 (baseline)	0.4108	21.72	68.56	53.69
Direct FT: $\mathcal{L}_{\text{disc}}$	**0.5294**	**67.14**	57.36	31.44
Direct FT: $\mathcal{L}_{\text{disc}} + \mathcal{L}_{\text{anchor}}$	0.4540	63.19	71.70	54.80
w/o multi-view context	0.4535	62.23	70.51	56.71
w/o teacher $\mathcal{L}_{\text{anchor}}$	0.4619	65.14	73.45	54.93
w/o student $\mathcal{L}_{\text{disc}}$	0.4350	44.88	72.51	56.54
Ours (full)	0.4632	66.01	**75.46**	**57.00**

Table 4: Ablation study. Sep./Corr. denote ScanNet feature separation and correspondence PCK@10px; Sem./Seg. denote SPair-71k semantic correspondence PCK@0.1 and ScanNet segmentation mIoU.

degrades performance across all four metrics. Removing the student discriminative objective causes a large drop in geometric correspondence, while removing the teacher anchor reduces both geometric and semantic performance. Removing multi-view context consistently reduces performance, confirming the benefit of cross-view geometric conditioning in the teacher. Overall, the full model achieves the most favorable balance between geometric discriminability and semantic transfer, retaining strong separation and correspondence while achieving the best semantic correspondence and segmentation among the ablated variants. Further ablations are provided in the supplementary appendix, including evaluations of off-the-shelf 3D-aware foundation features and different base models.

5 Conclusion

We presented DDMS, a 3D-aware feature distillation framework that transfers multi-view geometric knowledge into a single-view visual encoder. DDMS constructs a multi-view teacher through geometry-aware feature fusion, discriminative ranking, and semantic anchoring, and distills the resulting representation into a practical single-view student. Across correspondence, dense prediction transfer, and explicit 3D lifting evaluations, DDMS improves cross-view consistency and local discriminability while preserving the semantic utility of the original foundation features. Our results demonstrate that intermediate representations from multi-view geometry models provide effective supervision for learning stronger general-purpose 3D-aware visual features. Despite requiring only a single image at inference time, DDMS relies on posed multi-view RGB-D data and a geometry foundation model during training. Moreover, because the distilled student does not perform explicit multi-view reasoning at test time, it may remain limited when geometry is inherently ambiguous from a single observation. Exploring weaker geometric supervision and test-time multi-view extensions provides promising directions for future work.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Sony Faculty Innovation Award; the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and Alliance Advantage Grant; the Digital Research Alliance of Canada; and Advanced Research Computing at the University of British Columbia.

References

[1] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel. Deep ViT features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021.

[2] M. B. Azhari and D. H. Shim. DINO-VO: A feature-based visual odometry leveraging a visual foundation model. *IEEE Robotics and Automation Letters*, 2025.

[3] F. Baradel, M. Armando, S. Galaaoui, R. Brégier, P. Weinzaepfel, G. Rogez, and T. Lucas. Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision (ECCV)*, 2024.

[4] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[5] S. F. Bhat, I. Alhashim, and P. Wonka. AdaBins: Depth estimation using adaptive bins. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] J. Bratulić, S. Mittal, T. Brox, and C. Rupprecht. On geometric understanding and learned priors in feed-forward 3D reconstruction models. *arXiv preprint arXiv:2512.11508*, 2026.

[7] A. Brown, W. Xie, V. Kalogeiton, and A. Zisserman. Smooth-AP: Smoothing the path towards large-scale image retrieval. In *European Conference on Computer Vision (ECCV)*. Springer, 2020.

[8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021.

[9] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefeys, A. Geiger, T.-J. Cham, and J. Cai. MVSplat: Efficient 3D Gaussian Splatting from sparse multi-view images. In *European Conference on Computer Vision (ECCV)*, 2024.

[10] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

[12] X. Dong, J. Bao, Y. Zheng, T. Zhang, D. Chen, H. Yang, M. Zeng, W. Zhang, L. Yuan, D. Chen, et al. MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[13] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg. RoMa: Robust dense feature matching. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[14] J. Edstedt, D. Nordström, Y. Zhang, G. Bökman, J. Astermark, V. Larsson, A. Heyden, F. Kahl, M. Wadenbäck, and M. Felsberg. RoMa v2: Harder better faster denser feature matching. *arXiv preprint arXiv:2511.15706*, 2025.

[15] M. El Banani, A. Raj, K.-K. Maninis, A. Kar, Y. Li, M. Rubinstein, D. Sun, L. Guibas, J. Johnson, and V. Jampani. Probing the 3D awareness of visual foundation models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[16] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. In *International Journal of Computer Vision (IJCV)*, 2010.

[17] S. Fu, M. Hamilton, L. Brandt, A. Feldman, Z. Zhang, and W. T. Freeman. FeatUP: A model-agnostic framework for features at any resolution. In *International Conference on Learning Representations (ICLR)*, 2024.

[18] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *The international journal of robotics research*, 2013.

[19] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.

[20] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[21] J. Hou, S. Xie, B. Graham, A. Dai, and M. Nießner. Pri3D: Can 3d priors help 2d representation learning? In *International Conference on Computer Vision (ICCV)*, 2021.

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[23] R. Huang and K. Mikolajczyk. No Pose at All: Self-supervised pose-free 3D Gaussian Splatting from sparse views. In *International Conference on Computer Vision (ICCV)*, 2025.

[24] V. Jampani, K.-K. Maninis, A. Engelhardt, A. Karpur, K. Truong, K. Sargent, S. Popov, A. Araujo, R. Martin Brualla, K. Patel, et al. NAVI: Category-agnostic image collections with high-quality 3D shape and pose annotations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[25] N. Keetha, N. Müller, J. Schönberger, L. Porzi, Y. Zhang, T. Fischer, A. Knapitsch, D. Zauss, E. Weber, N. Antunes, et al. MapAnything: Universal feed-forward metric 3D reconstruction. *arXiv preprint arXiv:2509.13414*, 2025.

[26] B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis, et al. 3D Gaussian Splatting for real-time radiance field rendering. In *ACM Transactions on Graphics (TOG)*, 2023.

[27] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik. LERF: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023.

[28] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment Anything. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[29] Q. Li, J. Sun, L. An, Z. Su, H. Zhang, and Y. Liu. SemanticSplat: Feed-forward 3D scene understanding with language-aware Gaussian fields. *arXiv preprint arXiv:2506.09565*, 2025.

[30] H. Lin, S. Chen, J. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026.

[31] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys. LightGlue: Local feature matching at light speed. In *International Conference on Computer Vision (ICCV)*, 2023.

[32] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[33] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision (IJCV)*, 2004.

[34] Z. Luo, T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang, and L. Quan. GeoDesc: Learning local descriptors by integrating geometry constraints. In *European Conference on Computer Vision (ECCV)*, 2018.

[35] R. Mao, C. Bai, Y. An, F. Zhu, and C. Lu. 3DG-STFM: 3d geometric guided student-teacher feature matching. In *European Conference on Computer Vision (ECCV)*, 2022.

[36] J. Marrie, R. Ménégaux, M. Arbel, D. Larlus, and J. Mairal. LUDVIG: Learning-free uplifting of 2D visual features to Gaussian Splatting scenes. In *International Conference on Computer Vision (ICCV)*, 2025.

[37] J. Min, J. Lee, J. Ponce, and M. Cho. SPair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019.

[38] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[39] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. ORB-SLAM: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 2015.

[40] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[41] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser, et al. OpenScene: 3d scene understanding with open vocabularies. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[42] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[43] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister. LangSplat: 3D language Gaussian Splatting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Learning Representations (ICLR)*, 2021.

[45] M. Ranzinger, G. Heinrich, J. Kautz, and P. Molchanov. AM-RADIO: Agglomerative vision foundation model reduces all domains into one. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[46] M. B. Sarıyıldız, P. Weinzaepfel, T. Lucas, P. De Jorge, D. Larlus, and Y. Kalantidis. DUNE: Distilling a universal encoder from heterogeneous 2D and 3D teachers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[47] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[48] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[49] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[50] D. Shavin and S. Benaim. Splat and Distill: Augmenting teachers with feed-forward 3D reconstruction for 3D-aware distillation. In *International Conference on Learning Representations (ICLR)*, 2026.

[51] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[52] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGB-D images. In *European Conference on Computer Vision (ECCV)*, 2012.

[53] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

[54] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou. LoFTR: Detector-free local feature matching with transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[55] X. Sun, H. Jiang, L. Liu, S. Nam, G. Kang, X. Wang, W. Sui, Z. Su, W. Liu, X. Wang, et al. Uni3R: Unified 3D reconstruction and semantic understanding via generalizable Gaussian Splatting from unposed multi-view images. *arXiv preprint arXiv:2508.03643*, 2025.

[56] Z. Teed and J. Deng. DROID-SLAM: Deep visual slam for monocular, stereo, and rgb-d cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[57] H. Touvron, M. Cord, and H. Jégou. DeiT III: Revenge of the ViT. In *European Conference on Computer Vision (ECCV)*, 2022.

[58] V. Tschernezki, I. Laina, D. Larlus, and A. Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2D image representations. In *International Conference on 3D Vision (3DV)*, 2022.

[59] N. Tumanyan, A. Singer, S. Bagon, and T. Dekel. DINO-Tracker: Taming DINO for self-supervised point tracking in a single video. In *European Conference on Computer Vision (ECCV)*, 2024.

[60] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny. VGGT: Visual geometry grounded transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[61] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. DUSt3R: Geometric 3D vision made easy. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[62] Y. Wang, J. Zhou, H. Zhu, W. Chang, Y. Zhou, Z. Li, J. Chen, J. Pang, C. Shen, and T. He. Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025.

[63] P. Weinzaepfel, V. Leroy, T. Lucas, R. Brégier, Y. Cabon, V. Arora, L. Antsfeld, B. Chidlovskii, G. Csurka, and J. Revaud. CroCo: Self-supervised pre-training for 3d vision tasks by cross-view completion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[64] P. Weinzaepfel, T. Lucas, V. Leroy, Y. Cabon, V. Arora, R. Brégier, G. Csurka, L. Antsfeld, B. Chidlovskii, and J. Revaud. CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *International Conference on Computer Vision (ICCV)*, 2023.

[65] H. Xu, S. Peng, F. Wang, H. Blum, D. Barath, A. Geiger, and M. Pollefeys. DepthSplat: Connecting Gaussian Splatting and depth. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[66] S. Ye, Z.-H. Dong, R. Fan, T. Lv, and Y.-J. Liu. SemGS: Feed-forward semantic 3D Gaussian Splatting from sparse views for generalizable scene understanding. *arXiv preprint arXiv:2603.02548*, 2026.

[67] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *International Conference on Computer Vision (ICCV)*, 2023.

[68] Y. You, Y. Li, C. Deng, Y. Wang, and L. Guibas. Multiview equivariance improves 3D correspondence understanding with minimal feature finetuning. In *International Conference on Learning Representations (ICLR)*, 2025.

[69] Y. Yue, A. Das, F. Engelmann, S. Tang, and J. E. Lenssen. Improving 2D feature representations by 3D-aware fine-tuning. In *European Conference on Computer Vision (ECCV)*, 2024.

[70] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ADE20K dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[71] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong. iBOT: Image BERT pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

[72] S. Zhou, H. Chang, S. Jiang, Z. Fan, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, and A. Kadambi. Feature 3DGS: Supercharging 3D Gaussian Splatting to enable distilled feature fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

A Additional Analysis

We provide additional analyses to better understand the behavior and generality of the learned representation. These analyses examine the distribution of feature separation, geometry feature selection, off-the-shelf 3D-aware foundation features, backbone generalization, and out-of-distribution transfer.

A.1 Feature Separation Distribution

We further analyze the feature separation results reported in Table 1. While the main table reports a single average margin over all ScanNet [10] test pairs from the SuperGlue [47] protocol, this aggregate score does not show how separation varies across individual queries. We therefore analyze the per-query separation margins, using the same evaluation protocol, test pairs, query points, and negatives as in Table 1. Each sample is the positive-negative margin of one valid query point, pooled over all 1,500 test pairs; the mean therefore equals the separation reported in Table 1. Table 5 reports the mean and standard deviation of these per-query margins, while Fig. 9 visualizes the full distribution.

Method	Mean	Std.
DINOv2 [40]	0.4108	0.1935
Fit3D [69]	0.3494	0.1276
SnD [50]	0.4520	0.1388
SnD [50] (+blending)	0.3526	0.1233
MEF [68]	0.4288	0.1709
Ours	**0.4632**	**0.1226**

Table 5: Mean and standard deviation of the per-query feature separation margins on the ScanNet test pairs.

Fit3D and SnD (+blending) produce relatively compact distributions, but their margins are centered at lower values. DINOv2 and MEF achieve higher mean margins than these two, but show a much wider spread, with a heavier tail of low-margin queries. SnD without blending reaches a higher mean margin than DINOv2 and MEF with a moderate spread. In contrast, our method shifts the distribution toward larger margins and achieves the highest mean margin while maintaining the lowest standard deviation. This indicates that the proposed distillation improves feature separation consistently across queries, rather than through a small subset of highly distinctive points, supporting the intended balance between cross-view consistency and local discriminability.

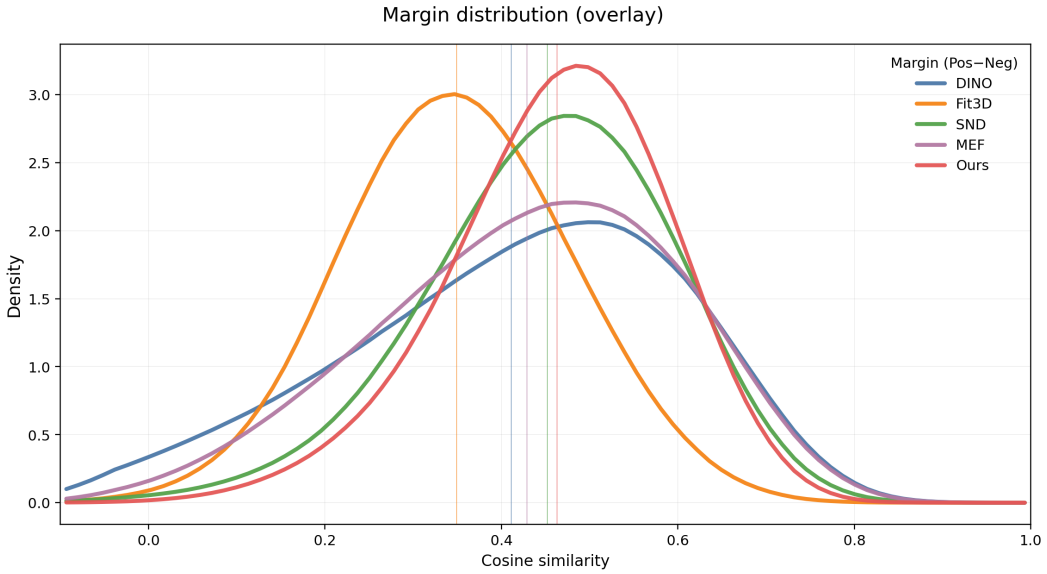

Figure 9: Distribution of per-query feature separation margins on the ScanNet [10] test pairs. Vertical lines mark the mean of each method.

Block	Stage	Sep. ↑	Corr. ↑
13	Early	0.3302	27.94
19	Middle	**0.3682**	**37.03**
39	Final	0.0860	13.54

Table 6: DA3 feature layer analysis. Frozen DA3-Giant features from early, middle, and final encoder blocks, evaluated directly on ScanNet feature separation and correspondence (PCK@10px).

Model	Scale	Params	Dim.	Mode	Sep. ↑	Corr. ↑	Sem. ↑	Seg. ↑
VGGT [60]	ViT-L + Agg.	1.26B	2048	Multi	0.3096	36.98	65.89	53.26
DA3 [30]	ViT-G	1.14B	1536	Multi	0.3682	37.03	48.74	39.90
DUNE [46]	ViT-B	86M	768	Single	0.1846	33.10	59.52	47.76
Ours	ViT-B	92M	768	Single	**0.4632**	**66.01**	**75.46**	**57.00**

Table 7: Evaluation of off-the-shelf 3D-aware foundation features. All features are evaluated frozen; Mode indicates multi-view (Multi) or single-view (Single) inference.

A.2 Choice of geometry feature layer

Our teacher uses intermediate features from Depth-Anything-3 (DA3) [30], a frozen geometry foundation model, as geometric context for refining pretrained 2D foundation features. We therefore compare features from early, middle, and final stages of DA3's cross-view reasoning process, evaluating the frozen features directly with the separation and correspondence protocol of Table 1. Specifically, block 13 corresponds to the first layer where cross-view attention is applied, block 19 represents an intermediate cross-view feature, and block 39 is the final layer of the DA3 encoder. As shown in Table 6, the middle-stage feature gives the best separation and correspondence performance. This suggests that teacher construction benefits from features that have undergone sufficient cross-view interaction, but are not yet overly specialized to the final geometry prediction head. Early features provide weaker geometric context, likely because they have only begun cross-view reasoning, while final features perform poorly as dense descriptors due to their specialization for the output geometry task.

A.3 Evaluation of Off-the-Shelf 3D-Aware Foundation Features

We evaluate whether existing 3D-aware foundation features can directly serve as general-purpose dense visual representations. For VGGT [60] and Depth-Anything-3 (DA3) [30], we use their commonly adopted large-scale variants and evaluate the frozen multi-view features directly. DUNE [46] is a single-view universal encoder distilled from heterogeneous 2D and 3D teachers [40, 61, 3], and we evaluate its frozen dense representation. This comparison evaluates whether strong off-the-shelf 3D-aware representations are already sufficient as dense visual features. We report ScanNet feature separation (Sep.) and correspondence PCK@10px (Corr.), SPair-71k semantic correspondence PCK@0.1 (Sem.), and ScanNet semantic segmentation mIoU (Seg.). As shown in Table 7, off-the-shelf 3D-aware features are not automatically strong general-purpose dense features. Although VGGT and DA3 are much larger geometry foundation models with multi-view inference, DDMS achieves the strongest performance on all four metrics with a lighter single-view encoder. Compared with DUNE, which has a comparable parameter count and single-view inference mode, DDMS also shows stronger feature performance across separation, correspondence, semantic matching, and segmentation. These results suggest that geometry-aware features are most effective when used as context for refining and distilling pretrained 2D foundation features, rather than being used directly as the final representation.

A.4 Downstream Camera Pose Estimation

We further evaluate whether the improved feature correspondence transfers to downstream camera pose estimation on ScanNet [10], 7-Scenes [51], and ETH-3D [49]. For each dataset, evaluation pairs

Feature	ScanNet	7Scenes	ETH3D
DINOv2 [40]	15.1 / 26.9	76.1 / 90.3	16.7 / 36.5
Fit3D [69]	19.0 / 38.0	77.3 / 92.9	16.0 / 38.8
MEF [68]	30.7 / 48.1	88.3 / 98.4	26.2 / 54.0
SnD [50]	20.6 / 36.9	81.1 / 92.3	15.2 / 36.9
SnD [50] (+blending)	18.8 / 34.1	77.0 / 93.0	15.2 / 38.8
Ours	**65.9 / 85.4**	**95.4 / 100.0**	**49.0 / 80.6**

Table 8: Downstream camera pose estimation with PnP-RANSAC. Each entry reports pose recall R@10 / R@25 (%), within 10 cm/10° and 25 cm/10°.

Backbone	Model	Sep. ↑	Corr. ↑	Sem. ↑	Seg. ↑
ViT-L	DINOv2-L-reg	0.4315	23.68	67.83	55.84
	+ Ours	**0.5413**	**61.39**	**77.75**	**57.91**
ViT-B	DINOv2-B-reg	0.4167	26.61	68.21	54.67
	+ Ours	**0.5127**	**60.74**	**75.72**	**55.72**
ViT-B	DINOv3-B	0.3251	30.26	69.93	**57.10**
	+ Ours	**0.4455**	**64.76**	**73.85**	56.92

Table 9: Generalization across foundation backbones. Each backbone is compared with DDMS trained on top of it (+ Ours), using the metrics of Table 7.

are fixed across methods, and query poses are estimated using the same mutual nearest-neighbor matching and PnP-RANSAC pipeline. Ground-truth depth is used to construct the reference 3D points so that the comparison isolates descriptor quality. We report pose recall R@10 and R@25, the percentage of queries localized within 10 cm/10° and 25 cm/10° of the ground-truth pose. As shown in Table 8, DDMS achieves the highest recall on all three datasets, improving substantially over both the original DINOv2 features and the 3D-aware baselines, showing that the learned representation transfers from feature-level correspondence to a downstream geometric task.

A.5 Generalization Across Backbones

The main paper focuses on DINOv2 [40] ViT-B and ViT-S for fair comparison with competitive baselines that provide public checkpoints in these settings. We further evaluate DDMS with additional foundation backbones: DINOv2-L and DINOv2-B with registers [11] (DINOv2-L-reg, DINOv2-B-reg), and DINOv3-B [53]. These experiments test whether the proposed 3D-aware distillation strategy generalizes across backbone scales and pretraining variants. As in Table 7, Sep. and Corr. denote ScanNet feature separation and correspondence PCK@10px, and Sem. and Seg. denote SPair-71k PCK@0.1 and ScanNet segmentation mIoU. As shown in Table 9, DDMS consistently improves feature separation, geometric correspondence, and semantic correspondence across all tested backbones. The gains are especially pronounced for geometric correspondence, indicating that the proposed distillation strategy is not tied to a specific DINOv2-B/S configuration. For semantic segmentation, DDMS yields clear improvements on the DINOv2 variants and remains comparable on DINOv3, where the change is marginal. Overall, these results suggest that DDMS enhances 3D-aware feature quality across different foundation backbones while largely preserving dense semantic transfer.

A.6 Out-of-Distribution Dense Prediction Transfer

Fig. 10 shows qualitative examples from additional dense prediction datasets beyond the indoor RGB-D domains used in the main transfer evaluation. While the main dense prediction results focus primarily on indoor scene understanding, we further evaluate frozen-feature linear probes on ADE20K [70] and Pascal VOC [16] for semantic segmentation, and KITTI [18] for outdoor depth estimation, using the probe protocol of the main paper (KITTI images are processed at 1008×294 to retain sufficient vertical patch resolution). These datasets cover broader scene categories, object-centric images, and outdoor driving scenes, testing whether DDMS preserves general-purpose

	ADE-20K [70]			Pascal-VOC [16]			KITTI [18]		
Model	mIoU ↑	mAcc ↑	aAcc. ↑	mIoU ↑	mAcc ↑	aAcc. ↑	RMSE ↓	AbsRel ↓	δ_1 ↑
DINOv2	44.29	55.45	79.67	81.30	87.52	95.85	4.0027	0.1228	85.90
Ours	**47.14**	**57.81**	**82.15**	**84.97**	**90.16**	**96.82**	**3.7735**	**0.1207**	**86.17**

Table 10: Out-of-distribution dense prediction transfer. We evaluate frozen-feature linear probes on ADE-20K [70] and Pascal-VOC [16] for semantic segmentation, and on KITTI [18] for depth estimation.

transferability beyond the indoor domain. The quantitative results in Table 10 show that DDMS continues to perform well in these out-of-distribution settings.

B Additional Qualitative Results

We provide additional qualitative examples from the evaluation settings used in the main paper. These examples complement the quantitative results and illustrate the behavior of the learned representation across correspondence, dense prediction, and 3D aggregation tasks.

Geometric correspondence. Fig. 11 provides additional nearest-neighbor matching examples on ScanNet [10]. These examples evaluate feature matching between different views of the same indoor scene, where correct matches require geometric consistency and local discriminability.

In-the-wild correspondence. Fig. 12 provides additional nearest-neighbor matching examples on Navi-Wild [24]. Unlike ScanNet, these pairs contain the same object under different backgrounds, viewpoints, and visual contexts, testing whether the learned features remain robust beyond shared-scene geometry.

Semantic keypoint matching. Fig. 13 provides additional semantic keypoint matching examples on SPair-71k [37]. This benchmark matches keypoints across different object instances of the same category, testing whether the learned features preserve semantic correspondence ability beyond shared-scene geometry.

Semantic segmentation. Fig. 14 provides additional semantic segmentation examples from frozen-feature linear probing. These examples evaluate whether the learned representation preserves dense semantic structure for downstream transfer beyond correspondence-based tasks.

Depth estimation. Fig. 15 provides additional depth estimation examples from frozen-feature linear probing. We compare DDMS with the DINOv2 baseline to examine whether the learned representation preserves dense geometric cues under single-view inference. The examples show that DDMS better preserves scene layout and object-level depth structure.

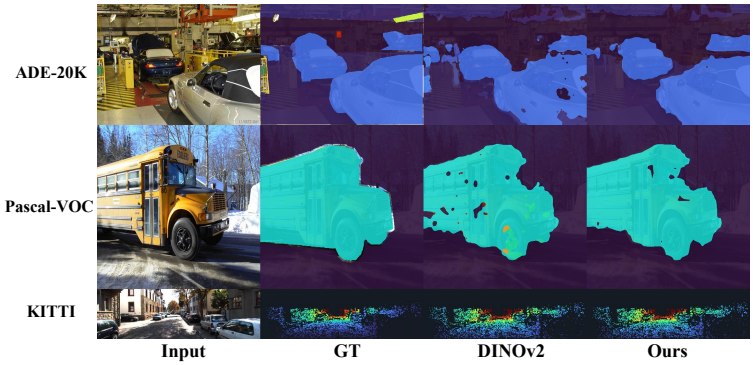

Figure 10: Qualitative out-of-distribution transfer with frozen-feature linear probes: semantic segmentation on ADE20K [70] and Pascal VOC [16], and depth estimation on KITTI [18].

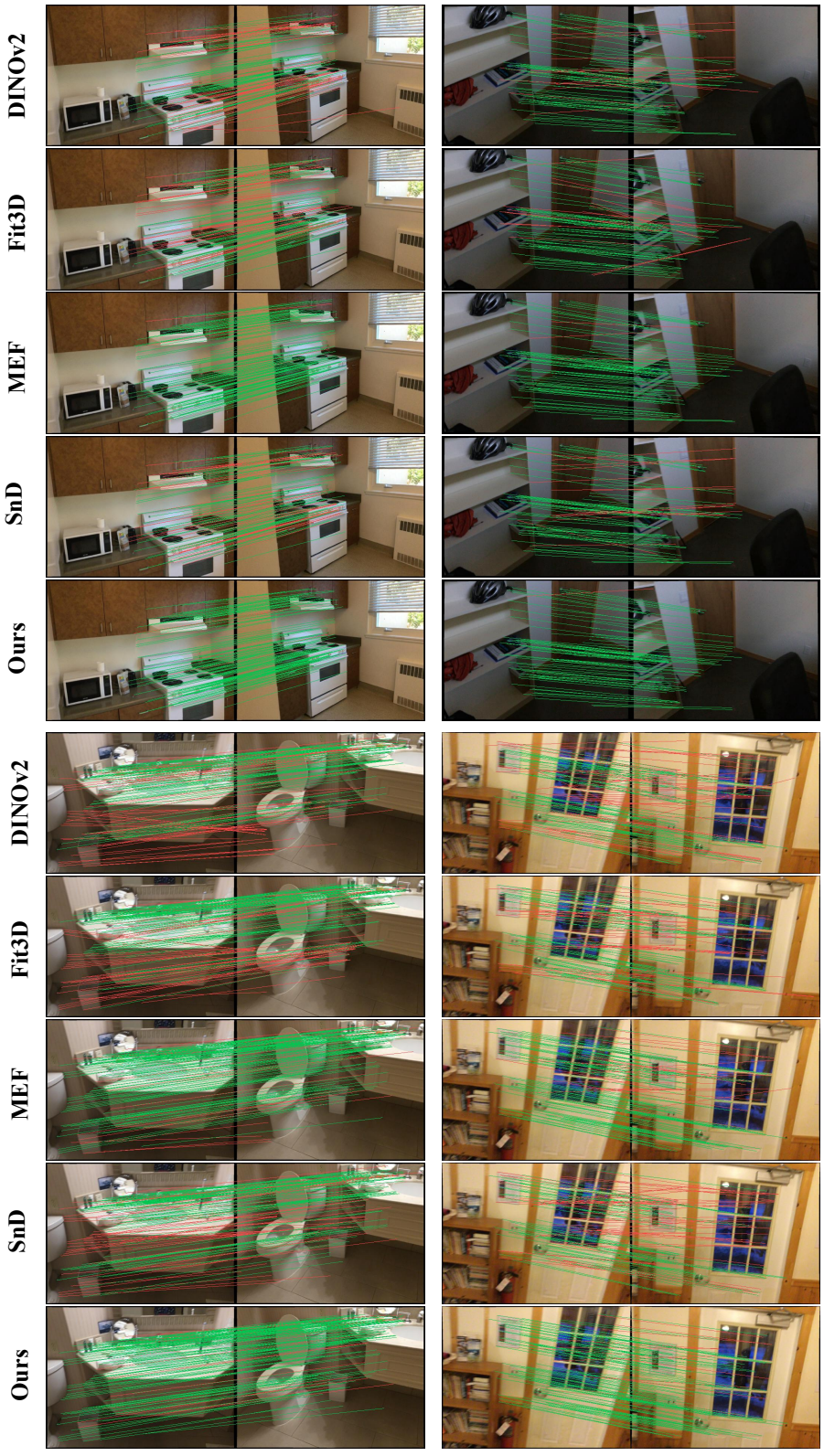

Figure 11: Additional nearest-neighbor correspondence results on ScanNet [10], compared with all baselines. Green and red lines indicate correct and incorrect matches.

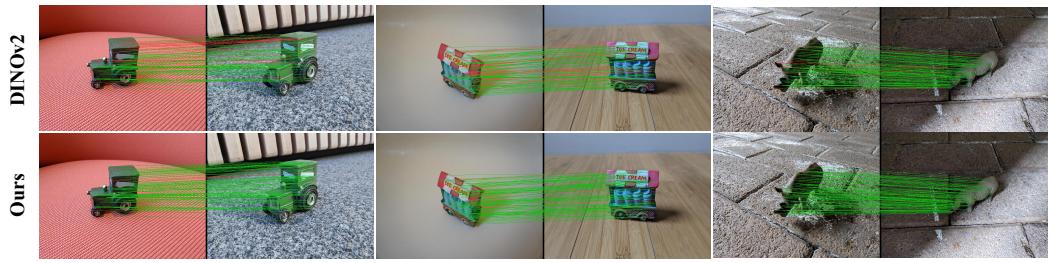

Figure 12: Additional correspondence results on Navi-Wild [24], where the same object appears under different backgrounds and viewpoints. DDMS is compared with DINOv2.



Figure 13: Additional semantic keypoint matching results on SPair-71k [37], where keypoints are matched across different instances of the same category.

3D point segmentation. Fig. 16 provides additional qualitative results for point-based 3D feature aggregation. After image features are lifted to 3D point clouds and evaluated with a linear probe, the results show whether the representation retains semantic structure after cross-view accumulation.

Feature splatting. Fig. 17 provides additional qualitative examples of feature splatting on pretrained 3D Gaussian scenes. Image features are lifted to the Gaussians using the same fixed uplifting procedure as in the main paper and rendered from held-out novel views, testing whether the feature representation remains coherent after 3D aggregation and rendering.

C Implementation Details

Training setup. We train DDMS on posed multi-view RGB-D sequences from ScanNet++ [67], using only training scenes and excluding evaluation scenes at the scene level. Camera poses and depth maps are used only for constructing geometric supervision and candidate labels. The final distilled student operates on a single image and does not require camera poses, depth maps, multi-view inputs, or the geometry model at inference time.

Teacher construction. We construct the multi-view teacher by fusing features from a pretrained 2D backbone with features from a frozen Depth-Anything-3 (DA3) model [30]. Unless otherwise specified, we use the feature from the 19th DA3-Giant encoder block, following the layer analysis in Table 6. DA3 remains frozen throughout training and provides multi-view geometric context through its intermediate features. For all teacher inputs, images are resized following the DA3 preprocessing, with the longer side set to 504 pixels while preserving the aspect ratio. The 2D and geometry features are concatenated, projected with a 1×1 convolution, refined with a 3×3 convolution, and added back to the original 2D feature as a residual update. This lightweight residual refinement head keeps the teacher representation in the original backbone feature space while injecting multi-view geometric

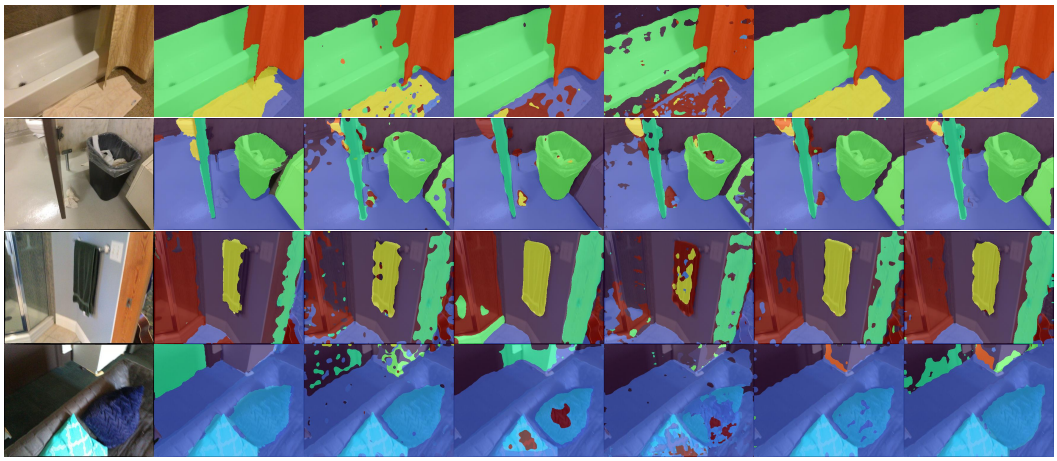

Figure 14: Additional semantic segmentation results from frozen-feature linear probing. All methods use the same linear probe.

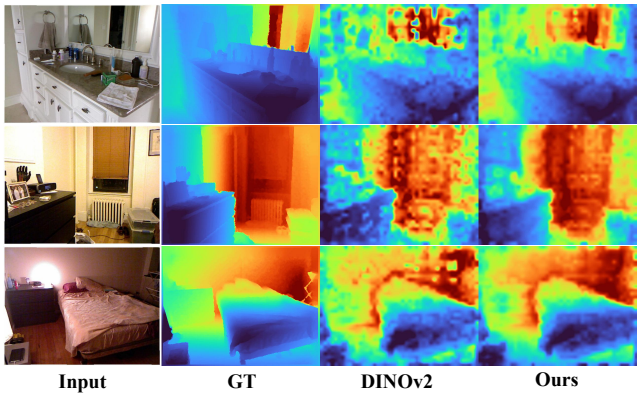

Figure 15: Additional depth estimation results from frozen-feature linear probing, comparing DDMS with DINOv2.

context. We train the teacher refinement stage for 20k steps using AdamW [32] with a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and gradient clipping with a maximum norm of 1.0. Each iteration uses two multi-view samples, each with $N = 4$ input views jointly processed by DA3. For the ranking objective in Section 3, we sample $M = 2$ views from the same scene and construct dense query-candidate matches. The trainable parameters are the fusion module and LoRA [22] adapters on the 2D backbone. For each query pixel, positives and negatives are assigned from a pool of 1,024 candidate pixels. Candidates within 0.05m are treated as positives, candidates farther than 0.15m as negatives, and candidates between the thresholds are ignored. We further filter candidates using reprojection validity and visibility/depth consistency. For the ranking loss, we use temperature 0.01. For the semantic anchor, we set $\lambda_{\text{anchor}} = 0.2$ and cosine-distance tolerance $\delta = 0.2$.

Student distillation. After teacher refinement, we distill the refined multi-view teacher into a single-view student. The student is initialized from the pretrained 2D foundation backbone and adapted with LoRA [22] of rank 32. The student also includes a lightweight 3×3 convolutional refinement head applied on top of the single-view backbone features. For each training sample, the frozen teacher processes $N = 4$ views jointly to produce refined multi-view teacher features. We select $M = 3$ views from this set, feed them independently to the student, and supervise the student with the objective in Equation (8), using $\lambda_{\text{disc}} = 0.2$. For the distillation loss, Ω_m denotes all valid patch locations in the feature map after image resizing; geometric validity filtering is used only for the ranking objective. One of the M views serves as the reference view: feature distillation

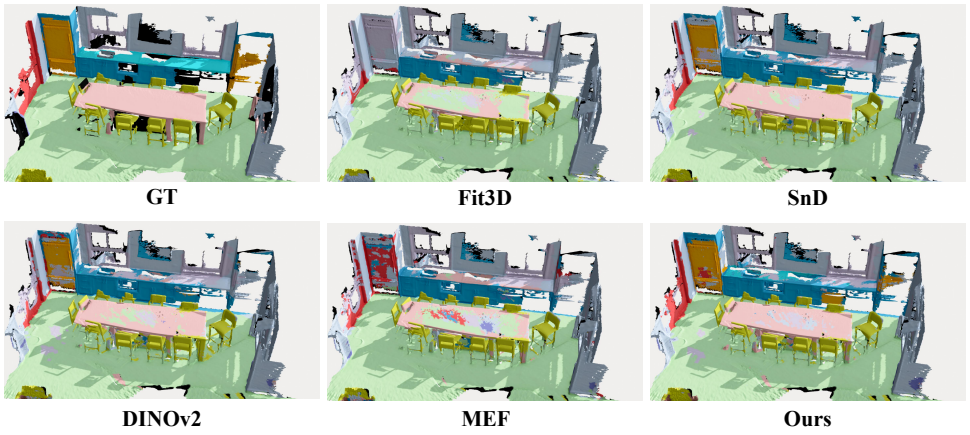

Figure 16: Additional 3D point segmentation results on ScanNet [10]. Features are aggregated from posed RGB-D views onto the point cloud and classified with a linear probe.

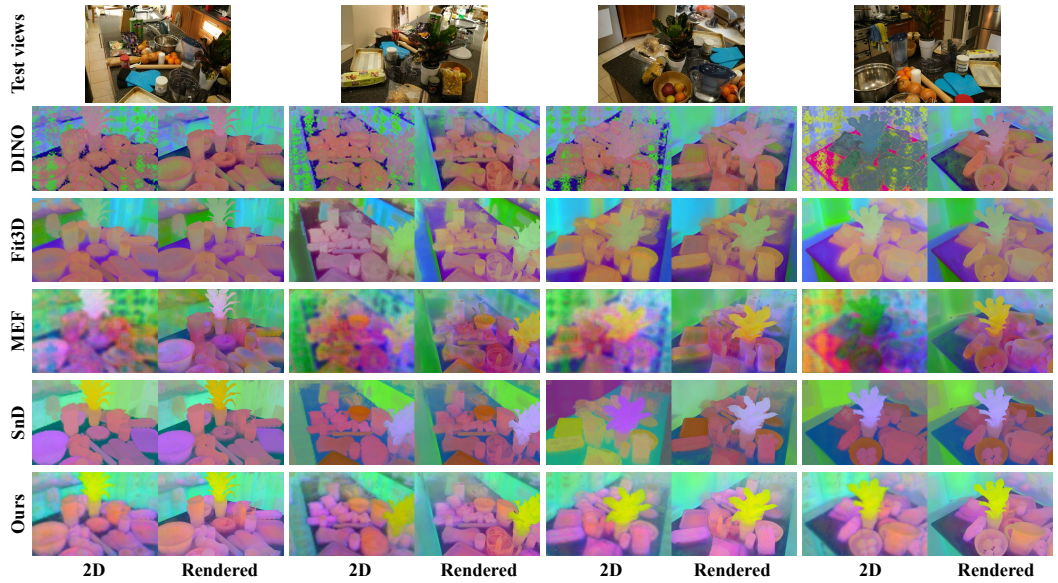

Figure 17: Additional feature splatting results on pretrained 3D Gaussian scenes. For each method, the original 2D features are shown next to the features rendered from held-out views after uplifting.

is applied to every view, and the discriminative loss to each pair formed by the reference view and one of the other views. For the student ranking loss, we use a temperature of 0.02, positive/negative 3D-distance thresholds of 0.05 m/0.10 m, and a 15-pixel exclusion radius around the ground-truth match for negatives, with 1,024 additional random candidate pixels sampled from the paired view. We tune the student for 20k steps using the same AdamW settings as teacher refinement with a cosine learning-rate schedule. At inference time, the student operates on a single image without requiring multi-view inputs, camera poses, depth maps, or the DA3 geometry model.

D Evaluation Protocols

All methods are evaluated with their publicly released weights and the same input resolution within each task (a 504-pixel longer side unless stated otherwise), using their own dense features without concatenating them with DINOv2 features. SnD [50] is reported with both of its released checkpoints: the checkpoint without SAM-mask blending is denoted SnD, and the checkpoint with blending is denoted SnD (+blending).

D.1 Geometric and Semantic Feature Quality

Feature separation. We evaluate feature separation on ScanNet [10] using the test image pairs from the SuperGlue [47] protocol. For each pair, we use the first image as the reference view and sample query pixels on a regular grid. Queries with invalid depth, invalid reprojection, or failed visibility/depth-consistency checks are discarded, and the same valid query set is used for all methods. For each valid query, we reproject the 3D point into the paired view using ground-truth depth and camera poses to obtain the positive correspondence. We then sample 50 random negative pixels in the paired view, excluding pixels within a 30-pixel radius of the positive location to avoid ambiguous nearby matches. The same query, positive, and negative locations are used for all methods. For each query, the separation margin is defined as the query-positive cosine similarity minus the average cosine similarity to the 50 negatives. We report the margin averaged over all valid queries.

Geometric correspondence. We evaluate geometric correspondence on ScanNet [10] and Navi-Wild [24]. For ScanNet, we use the test image pairs from the SuperGlue [47] protocol. Building on the protocol of Probe3D [15], we sample the dense features of each method on a common pixel grid at one quarter of the original image resolution (324×242), and match every grid point with valid depth in the first view to its nearest neighbor in the second view. Unlike Probe3D, we do not apply a ratio test or keep only the top-ranked matches, so all methods are scored on the same set of queries. We report PCK@10px and PCK@20px on this grid, with ground-truth correspondences obtained from depth and camera poses; Fig. 4 reports the same metric binned by the relative rotation between the two views.

For Navi-Wild, we evaluate correspondence under in-the-wild object-centric conditions, where the target object appears across different contexts, viewpoints, and visual conditions. We use image pairs from the Navi test set with relative viewpoint changes below 120° and sample query points only inside the provided object mask. We further discard queries whose ground-truth point is not visible in the target view, i.e., whose reprojection falls outside the target object mask or disagrees with the target depth by more than 5%. Following the Probe3D protocol [15], correspondences are obtained by nearest-neighbor search in feature space and evaluated using PCK@0.05 and PCK@0.10 with normalized image coordinates. This measures whether predicted correspondences fall within the specified normalized distance from the ground-truth match, testing robustness beyond shared-scene geometric matching.

Semantic correspondence. We evaluate semantic correspondence on SPair-71k [37] and DAVIS-2017 [42]. For SPair-71k, we use the official test set and restrict evaluation to image pairs with view-difference level 0. Given a source keypoint, we find its nearest neighbor in the target image in feature space and compare the predicted location with the annotated target keypoint. We report PCK at the normalized thresholds used in the main paper, measuring the percentage of predicted correspondences that fall within the specified distance from the ground-truth keypoint. This evaluates semantic matching across different object instances of the same category.

For DAVIS-2017, we evaluate temporal correspondence through video object mask propagation on the validation split. Given the ground-truth mask in the first frame, labels are propagated to target frames using frozen backbone features without task-specific training. Labels are transferred from memory-frame patches to target-frame patches based on feature similarity, and the predicted masks are evaluated with the standard DAVIS metrics: region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$. We report the average over two memory settings: first-frame-only propagation and propagation using the first frame plus the most recent past frame.

D.2 Dense Prediction Transfer

Semantic segmentation. We evaluate semantic segmentation with linear probing on frozen dense features. For each method, the backbone is fixed and only a one-layer spatial linear classifier is trained on top of the dense feature map. The classifier predicts semantic logits at the feature-grid resolution, which are bilinearly upsampled to the ground-truth label resolution for loss computation and evaluation. We train the probe with cross-entropy loss on valid labeled pixels, ignoring unlabeled or void regions. We evaluate on ScanNet [10] and ScanNet++ [67], using the official train/validation scene splits of both datasets. All methods use the same splits, probe architecture, optimizer settings,

and 50k-iteration training schedule. We report mean intersection-over-union (mIoU), mean class accuracy (mAcc), and overall pixel accuracy (aAcc).

Depth estimation. We evaluate depth estimation with linear probing on frozen dense features, following the setup used in our Fit3D and SnD comparisons. For each method, the backbone is fixed and only a one-layer depth classifier is trained. Following common bin-classification depth estimation protocols [5], we concatenate the global `[CLS]` token to each spatial patch token, predict a distribution over discretized depth bins, and bilinearly upsample the logits before aligning them with the ground-truth depth map. The probe is trained with cross-entropy loss on valid depth pixels within the dataset-specific depth range. At inference time, continuous depth is recovered as the expected depth under the predicted bin distribution. We evaluate on ScanNet [10] and NYUv2 [52], training on the ScanNet 25k-frame subset and evaluating on the official ScanNet test scenes, and using the standard NYUv2 train/test split. All methods use the same probe capacity, optimizer settings, depth-bin configuration, and 50k-iteration training schedule. We report AbsRel, RMSE, and δ_1, where δ_1 is the percentage of valid pixels satisfying $\max(\hat{d}/d, d/\hat{d}) < 1.25$.

D.3 3D Lifting and Rendering

3D feature uplifting and rendered matching. We evaluate whether the learned 2D features remain reliable after explicit lifting into 3D and rendering to novel views. We use LUDVIG [36], a learning-free feature uplifting method for Gaussian Splatting scenes. Given a pretrained Gaussian scene, LUDVIG assigns 2D image features to 3D Gaussians by aggregating features according to each pixel's rendering contribution. This contribution-weighted procedure avoids an additional feature-optimization stage, allowing us to evaluate the liftability of the frozen 2D representation itself. We use only the direct uplifting step and do not apply the graph-diffusion refinement from LUDVIG.

We evaluate on Mip-NeRF 360 scenes [4] using Gaussian Splatting models trained for 30,000 iterations with the standard setup [26]. Every eighth image is held out as a test view; the remaining images are used to train the Gaussian scenes and uplift features. Following the official LUDVIG implementation, we reduce all features to 40 dimensions with PCA before uplifting to make feature rendering tractable under a common pipeline. After uplifting, we render the 3D feature field from held-out test poses and evaluate nearest-neighbor correspondence under rendered-to-2D, 2D-to-rendered, and rendered-to-rendered settings. Matches are evaluated using PCK at the pixel thresholds reported in the main paper.

3D point segmentation. We evaluate 3D point segmentation with linear probing on ScanNet [10] point clouds. Following the feature aggregation strategy used in OpenScene [41], we lift 2D image features onto 3D points using the provided depth maps and camera poses. For each 3D point, we project it into posed RGB-D frames, sample the corresponding 2D feature at each valid projected pixel, and average the sampled features across valid views. The aggregated 3D point features are kept frozen, and only a linear classifier is trained for semantic segmentation. All methods use the same point clouds, camera poses, aggregation procedure, train/validation split, classifier capacity, and optimization settings. For the fused features, we report mIoU, mAcc, and aAcc. To expose disagreement that averaging can hide, we additionally apply the same classifier to the feature of every individual view in which a point is visible, without averaging. We report the mIoU over all point–view pairs (per-view mIoU) and the multi-view agreement, i.e., the probability that two randomly chosen views of a point predict the same class, averaged over all points visible in at least two views.