DINOCULAR: SELF-SUPERVISED VISUOSPATIAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a self-supervised framework for learning joint visuospatial representations from RGB-D observations. While modern vision foundation models are trained almost exclusively on RGB images, many embodied systems have access to explicit depth sensing, which provides geometric information that monocular inputs cannot recover. Our method integrates depth-derived geometric priors with a visual backbone through inter-patch and intra-patch fusion, enabling the model to encode both appearance and spatial structure efficiently. The resulting representation shows promising improvements on 3D awareness while preserving semantic transfer: it outperforms prior methods of comparable scale on multiple 3D geometry benchmarks, and remains competitive when probed for standard RGB-D semantic segmentation tasks.

1 INTRODUCTION

Explicit depth sensing is a remarkably widespread biological solution for spatial perception: binocular stereopsis, which compares observations from two eyes to infer depth, has been demonstrated across diverse animal species, from primates and other mammals (Heesy, 2009) to falcons (Fox et al., 1977) and cuttlefish (Feord et al., 2020). In artificial embodied systems, however, the dominant vision models remain largely monocular. Despite the availability of stereo cameras and depth sensors in robots, autonomous vehicles, mixed-reality headsets, and consumer smartphones, modern vision foundation models are still trained almost exclusively on RGB images. This mismatch has left frontier visual AI systems with a persistent weakness in spatial and 3D understanding (Feng et al., 2025; El Banani et al., 2024; Man et al., 2024; Azzolini et al., 2025). Training with RGB inputs on more spatial tasks and data cannot fully solve this: Without depth as an input modality, spatial reasoning remains physically constrained by the scale ambiguity of monocular sensing.

Depth has historically been difficult to incorporate into foundation-scale vision learning for several reasons. Compared with RGB images, depth data is scarce, heterogeneous, and noisy: it may come from stereo, structured light, time-of-flight sensors, or sparse LiDAR, each with different failure modes and sampling patterns. At the same time, the field has become increasingly reliant on large RGB-only feature extractors such as DINO, making it natural to treat geometry as an output to be predicted rather than as an input to be encoded. Recent progress in learned geometric reconstruction (Wang et al., 2025a; Keetha et al., 2026; Wang et al., 2024) changes this picture. We can generate depth for training at very large scales, and we can homogenize different depth sources through completion and denoising into dense pixel-wise depth maps. These advances reopen a central question: how can models learn general-purpose representations that jointly encode visual appearance and geometric structure?

Existing approaches to visuospatial learning fall broadly into two categories. The first post-trains RGB-only foundation models for geometric consistency or reconstruction. Breakthrough systems such as VGGT (Wang et al., 2025a) infer spatial structure from RGB observations, and recent work adapts visual representations to become more multi-view or spatially consistent. However, because these methods remain rooted in RGB-only inputs, they cannot fully exploit the complementary information provided by direct geometric sensing. The second category consists of task-specific RGB-D fusion architectures. Recent models have moved away from redundant dual-encoder designs toward more efficient mechanisms that use depth as a geometric prior. DFormerv2 (Yin et al., 2025), for example, introduces Geometry Self-Attention to guide feature interactions using depth while

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.

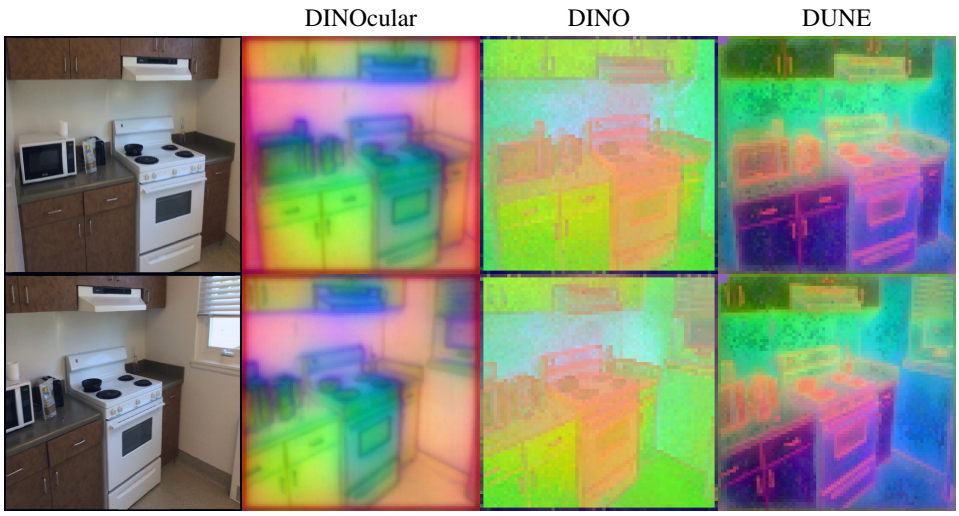

Figure 1: Visualization of the 3 strongest PCA components for DINOcular (ours), DINO ViT-B (Caron et al., 2021), and DUNE (Sarıyıldız et al., 2025). The mix of semantic and geometric information in our features is clearly visible at the edges of the kitchen objects.

avoiding the cost of a fully separate depth stream. Yet, our experiments reveal that these architectures do not automatically learn generally useful visuospatial representations.

In this paper, we propose a scalable, self-supervised framework for learning joint visuospatial representations from RGB-D observations. Our architecture efficiently integrates depth-derived geometric priors with a visual backbone, refining their interaction at inter-patch level through depth-aware 3D positional encoding and intra-patch through feature fusion. Coupled with a scalable, self-supervised training recipe over dense RGB-D observations, the resulting representation improves 3D awareness while preserving strong semantic transfer. Empirically, our model outperforms prior representations on multiple 3D benchmarks, including Probe3D (El Banani et al., 2024), and achieves competitive performance on standard RGB-D semantic segmentation benchmarks such as NYU DepthV2 (Silberman et al., 2012) and SUNRGBD (Song et al., 2015). Our main contributions are:

- We introduce an improved architecture that integrates visual and spatial information at both inter-patch and intra-patch levels.
- We present a self-supervised training recipe for learning generalizing visuospatial representations from RGB-D observations.
- We show that the learned representation improves over prior methods of comparable scale on 3D geometry benchmarks while remaining competitive on semantic segmentation tasks.

2 Related Work

2.1 Spatial Understanding in Vision Foundation Models

Vision Transformers (Dosovitskiy et al., 2021) established patch tokens as the common interface for self-supervised pretraining and downstream feature transfer. Two scalable families dominate the landscape. Self-distillation, where DINO (Caron et al., 2021) and iBOT (Zhou et al., 2021) were scaled to very large unlabeled corpora in DINOv2 (Oquab et al., 2023) and DINOv3 (Siméoni et al., 2025), yields some of the strongest available image features. Masked reconstruction (He et al., 2022a; Wang et al., 2023) forms the second family and targets pixel-level recovery. We adopt the self-distillation objective and extend it from RGB to RGB-D.

These features are strongly semantic but only weakly geometric. Probe3D (El Banani et al., 2024) and Lexicon3D (Man et al., 2024) report limited geometric consistency of frozen VFM features across viewpoints, and You et al. (2025) show that view consistency emerges only after explicit multi-view finetuning.

A complementary line learns geometry directly from images. CroCo (Weinzaepfel et al., 2022) introduced this thread with self-supervised cross-view completion, and its encoders underpin DUSt3R (Wang et al., 2024), MASt3R (Leroy et al., 2024), and VGGT (Wang et al., 2025a), which cast multi-view reconstruction as feed-forward pointmap and unified geometry prediction. Follow-up work generalises the paradigm to dynamic scenes, persistent state, and broad metric reconstruction (Zhang et al., 2025; Wang et al., 2025b; Keetha et al., 2026). Across this family RGB is the input and geometry is the output, so the feature space is shaped to predict 3D structure rather than to represent it.

Both families therefore leave depth outside the feature extractor, either absent in RGB-only SSL or present only as a prediction target. What is absent are generalizing representations that combine RGB and Depth inputs.

2.2 RGB-D Fusion and Geometric Priors

Early RGB-D models used dual-encoder architectures with convolutional or attention-based fusion (Hu et al., 2019; Chen et al., 2020), transformer-based models shifted fusion to the token level (Wang et al., 2022), and a more recent body of work refines supervised RGB-D segmentation through primary-modality distillation, heterogeneous branches, and robustness to degraded depth (Gong et al., 2026; Jaganathan & Vela, 2026). These methods train the encoder jointly with a segmentation head, so the representation inherits the inductive bias of the target task and does not transfer cleanly to other objectives (Yin et al., 2024).

DFormer (Yin et al., 2024) also targets RGB-D segmentation, but pretrains a backbone on image-depth pairs from ImageNet under supervised classification, while DFormerv2 (Yin et al., 2025) derives a geometry prior from pooled patch distances and relative depth, applied as an additive bias on self-attention weights. This prior operates at the pooled-patch level and modifies attention weights rather than token values or embeddings, so depth informs how patches attend but does not enter the feature representation itself. We instead embed depth at the level of token positions. Modern vision transformers encode token position relatively through RoPE (Su et al., 2024), and Schenck et al. (2025) extend the mechanism to a third axis for robotic action models, but whether the benefits persist for visuospatial representation learning is an open question. We follow that direction and treat image-plane coordinates together with mean patch depth as a joint position, so geometry shapes feature computation rather than reweighting a 2D affinity.

Several self-supervised methods incorporate geometry without producing a native RGB-D image backbone. MultiMAE (Bachmann et al., 2022) treats depth as a reconstruction target rather than as input geometry. Concerto (Zhang et al., 2026) jointly trains a 3D point cloud transformer and a 2D image encoder, but still requires a point cloud at inference and uses images only as a supervisory signal. Sonata (Wu et al., 2025a) trains a self-supervised point cloud encoder on aggregated scene-level scans and suppresses spatial information to avoid a geometric shortcut, where coordinate-based positional encodings cause representation collapse. In contrast, we target the more broadly available single-view RGB-D setting, avoiding assumptions about aggregated scans, fixed cameras, or scene-level point clouds.

3 Method

We investigate visuospatial grounding in two dimensions: model architecture and learning objectives. In the following, we first discuss how we integrate depth as an input modality to modern vision transformer architectures. Second, we describe the self-supervised learning objectives that we train these models on. Third, we describe how we modify a Swin-style architecture to implement these methods.

3.1 Geometry Prior

Inspired by the effective approach of incorporating depth in DFormerv2 (Yin et al., 2025), we explore approaches that make use of depth as a geometry prior, without creating large additional branches or complex fusion mechanisms. In particular, although Yin et al. (2025) achieves good performance, it is highly dependent on RMT (Fan et al., 2024), and simply biasing self-attention based on the

weighted depth distances. It uses depth to attend to parts of the input image that are metrically close by, strictly baking into the architecture the heuristic that close-by parts matter.

Instead of additive biasing, we extend rotary positional encoding (RoPE) (Su et al., 2024) into three dimensions. Given that RoPE is a foundational component in the self-attention mechanisms of most modern Vision Transformers (ViTs), extending it to a third dimension allows the model to inherently leverage relative spatial relationships based on the depth input. By encoding depth directly into the coordinate system (as 3rd, z-axis), we create a more principled prior that enables the model to reason about the physical proximity of features, but e.g. also to attend to patches further away. Following Schenck et al. (2025), for patch coordinates $u, v \in \mathbb{N}$ on the image plane and average depth $\bar{z} \in \mathbb{R}^{\geq 0}$ of all pixels within a patch, we find the encoding $\hat{x}$ of the patch embedding x as:

$$\hat{x} = \exp\left(\mathbf{L}_1 u + \mathbf{L}_2 v + \mathbf{L}_3 \bar{z}\right)\, x$$

where each $\mathbf{L}_k \in \mathbb{R}^{d \times d}$ is a rotational positional embedding $\mathbf{L}_k = \sum_{p=1}^{d/2} \left(\delta_{2p,2p-1} - \delta_{2p-1,2p}\right) \theta_p$ with learnable frequencies θ_p. Thus, we completely change approach seen in DFormerv2 with our inter-patch mechanism - 3D RoPE.

3D positional encoding of patches gives the model access to the global geometry of an observation. However, it does not make use of the pixel-level resolution of depth input that also holds important information of the local shape, e.g. whether edges are round or sharp or whether surfaces are uneven or smooth. We therefore add this local geometric information to the patchified RGB embeddings by extracting local features from the depth within each patch. We apply lightweight patch embedding layer and fuse both embeddings through a linear projection, giving the model intra-patch geometry information. One might suspect that scaled, metric distance is not relevant for local geometric details, and geometric intra-patch embedding would generalize better from a scale-free, normalized input. In our experiments, we therefore investigate an ablation where local shape features are instead computed based on per-pixel surface normals. Surface normals are the spatial derivative of the depth value and therefore scale-free. However, empirically we observe this variant to collapse in training.

3.2 Learning objective

We design the learning objective in a way that it should explicitly learn fine-grained spatial information along with the visual-semantic content. As semantic loss objective $\mathcal{L}_{sem}$, we utilize the teacher-student distillation framework of DINO and DINOv2 (Caron et al., 2021; Oquab et al., 2023). We briefly summarize how we apply these and refer the interested reader to Oquab et al. (2023) for more details:

- **DINO / Image-level objective** We give student and teacher different crops of the same object-centric image and take the `cls` tokens from each. They are passed through a DINO head and finally a softmax activation. $\mathcal{L}_{\text{DINO}}$ is then the cross-entropy between the student's output and the centered teacher's output.
- **iBOT / Patch-level objective** For the student, some of the input tokens are masked. Importantly, that means in our case that both RGB and intra-patch Depth are masked, but even the masked patch tokens are positionally encoded based on average patch depth. Patch tokens are then projected through an iBOT head and a softmax activation. $\mathcal{L}_{\text{iBOT}}$ is then cross-entropy between the student's output and the centered teacher's output over all masked patches.

In our experiments, DINO and iBOT heads share the same parameters, and we do not apply the untying of Oquab et al. (2023). That is because our experiments cannot reproduce the data scale at which they observed untying to be useful. However, we follow Oquab et al. (2023) to apply the KoLeo regularizer (Sablayrolles et al., 2018), arriving at the full semantic-visual objective

$$\mathcal{L}_{sem} = \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + \lambda \mathcal{L}_{\text{KoLeo}}$$

To fully leverage the spatial information of depth, we combine the above loss with a spatial objective. Our reasoning is that while visual appearance may change between viewpoints, the same part of an object is more easily matchable in 3D. Given the additional spatial input to the model, we require an objective that enforces the use of this information. We therefore experiment with different formulations of a multi-view consistency loss. In all cases, student receives two different images of the same object, but from different viewpoints. From known point maps and the depth, we find a

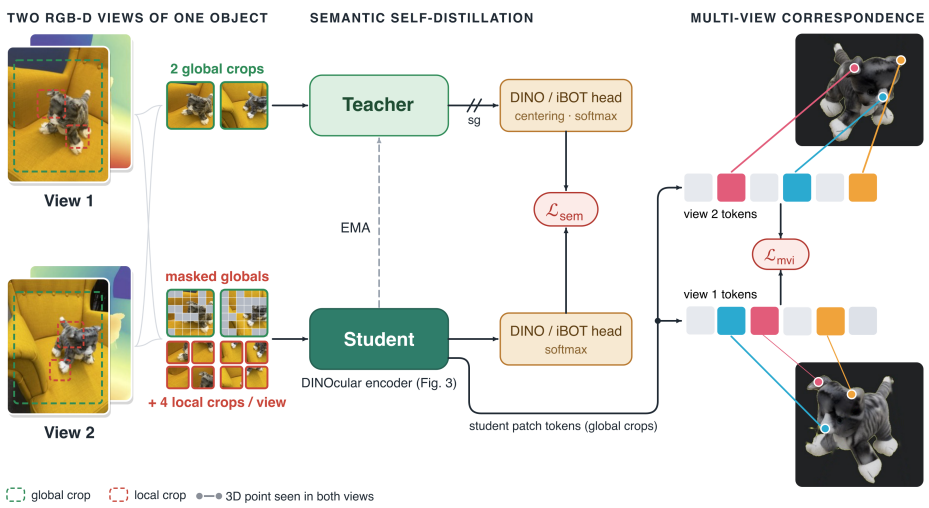

Figure 2: DINOcular multi-view pretraining setup: Model learns visual-semantic content via self-distillation ($\mathcal{L}_{sem}$) and spatial information via multi-view correspondence ($\mathcal{L}_{mvi}$).

subset of patches that is visible to both. We then enforce similarity between each pair of matched patches (z_1, z_2). This is visualized in Figure 2.

- **Cosine Similarity** Following Koch et al. (2025), who propose a post-training refinement of DINOv2 features with this objective, we optimize patch embeddings to have low cosine distance: $\mathcal{L}_{\text{mv-cosine}} = 1 - z_1 \cdot z_2$.
- **Contrastive Ranking** Following You et al. (2025), who investigate finetuning of DINOv2 and other vision foundation models for multi-view consistency, we apply a SmoothAP (Brown et al., 2020) ranking-based loss over the matched patches.
- **Multi-view iBOT** Since the iBOT objective works on a patch level, a variant of this objective for matched patches from different views is easily conceivable. In this variant, the second image however is passed through the teacher. We bias the mask sampling to those patches with visual overlap and retrieve pairs z_1 and z_2, where z_1 has a masked input. Both are passed through a projection head and softmax activation, and centering (for the teacher). $\mathcal{L}_{\text{mv-ibot}}$ is then the cross-entropy between these 2 outputs.

Our ablations find empirically that $\mathcal{L}_{\text{mv-contrastive}}$ is the best compatible with $\mathcal{L}_{sem}$, leading to our full objective $\mathcal{L} = \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + \lambda_1 \mathcal{L}_{\text{mv-contrastive}} + \lambda_2 \mathcal{L}_{\text{KoLeo}}$. We sample batches such that in every batch there are pairs of images to which the multi-view loss can be applied.

During model development, we observed that the combination of multi-view training and added intra-patch depth embedding yielded features that were heavily biased towards spatial tasks and contained less visual information. This hints at the connection between depth input and the multi-view consistency that we hypothesize above, but yields less general features. We address this by applying dropout to the intra-patch depth embedding layers, forcing the multi-view objective to not exclusively rely on depth information.

3.3 Technical Implementation

The backbone architecture is a Swin-style hierarchical vision transformer (Liu et al., 2021) with 4 stages presented in Fan et al. (2024) and further used in Yin et al. (2025), who added RGB-D blocks. We further completely replace mechanisms used in Yin et al. (2025) with our 3D RoPE discussed in 3.1 for inter-patch visuospatial information. And additionally provide first RGB-D block with intra-patch geometry as lightweight depth embedding, leading to an even more efficient architecture than DFormerv2 (Appendix B). The architecture is presented in Figure 3, and additional training details are reported in Appendix F.

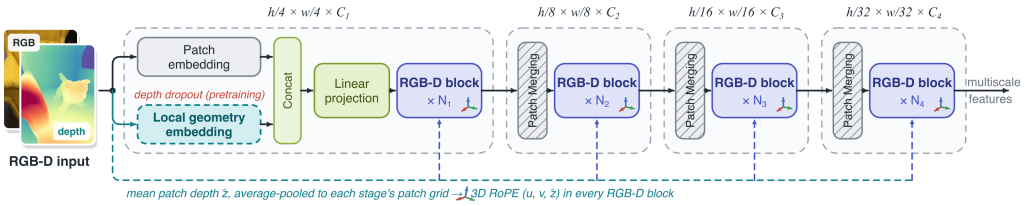

Figure 3: DINOcular encoder: Architecture with 3D RoPE and local geometry embedding.

4 Experiments

We empirically test our proposed method over a range of visuospatial tasks: RGB-D Semantic Segmentation, 3D Correspondence Estimation, and Object Pose Estimation. We first discuss dataset preparations and chosen methods of comparison. Then we justify our method choices in ablation studies. Finally, we compare our trained model with prior work on the different visuospatial tasks. We additionally report efficiency measurement in Appendix B and qualitative results in Appendix D.

4.1 Training Data and Details

We generate raw depth values for all our training samples using monocular depth estimation on ImageNet-1k and multi-view triangulation on MVImgNet2.0. For both, we use MapAnything (Keetha et al., 2026). This follows the approach of prior work (Yin et al., 2025; 2024), but uses a more accurate model. Our multi-view objective additionally requires point maps. We save point maps from MapAnything during depth generation. For ImageNet-1K, we use all $1.3\,\mathrm{M}$ images. For MVImgNet2.0, we keep all $307\,\mathrm{k}$ objects, but reduce the redundancy of many different viewpoints to approximately 8 per object.

4.2 Baselines

We compare against methods that differ in how, if at all, geometry enters the learned features. **DFormerv2** (Yin et al., 2025) takes RGB-D input and is our most direct comparison: we adapt a similar RGB backbone to the RGB-D setting, but it is trained with supervised classification. **MultiMAE** (Bachmann et al., 2022) is RGB-D backbone trained in a self-supervised manner, but uses depth as a masked reconstruction target rather than as a geometric prior only. **DUNE** (Sarıyıldız et al., 2025) takes RGB only and instead distills geometric information from MASt3R alongside DINOv2 features. The RGB-only **DINO family** (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025) provides the self-supervised objective we build on; DINO is trained on ImageNet-1K and directly comparable, whereas DINOv2 and DINOv3 use $\times 36$ and $\times 433$ more, non-public training data. We therefore additionally retrain DINOv2 in our setting on ImageNet-1K and on our full $3.9\,\mathrm{M}$ images, and evaluate public checkpoints at comparable data scale (Wu et al., 2025b). **Sonata** (Wu et al., 2025a) is also a self-supervised visuospatial encoder, but operates on merged scene-level point clouds, for which single-view evaluation is ill-posed. Large-scale DUNE and DINOv2/v3 models, and Sonata are reported for reference only; all remaining comparisons are at matched parameter count and training data. Finally, two ablations of our own model, without depth input and without the multi-view objective, isolate the contribution of each. Full baseline details are discussed in Appendix C.

4.3 Ablation: Model Architecture

We present an ablation of our architectural choices to encode inter- and intra-patch depth in Table 1. Surprisingly, even for the Dformerv2 architecture itself we find that exchanging the supervised with the DINO objective increases usefulness of the learned features with respect to segmentation probing, even though Yin et al. (2025) specifically target segmentation as their goal task. Our modification of the source RGB architecture with 3D RoPE instead of proximity attention bias yields the largest per-step increase in feature probing, for all 3 measured indicators. Contrary to the results from Tziafas & Kasaei (2023), the encoding of intra-patch geometry through surface normals does not show good

Table 1: **Ablation on geometry encoding.** We report RGB-D semantic segmentation probing (Segm.) on NYUDepthv2 (Silberman et al., 2012) (mIoU) and correspondence estimation with viewpoint change on NAVI (Jampani et al., 2023) (θ_{90}^{120}, recall) and ScanNet (θ_{60}^{180}, recall) following Probe3D (El Banani et al., 2024). *Proximity* is the geometry prior of DFormerv2 (Yin et al., 2025).

change	Attention	Intra-patch	DINO	iBOT	MVI	Dataset	Depth Dropout	Segm. [mIoU]	NAVI θ_{90}^{120} [R]	ScanNet θ_{60}^{180} [R]	
	Proximity	–	supervised			IN1k	–	24.86	17.40	7.30	
self-sup.	Proximity	–	✓			IN1k	–	25.78	16.40	4.50	**attention/ inter-patch** geometry
+2D RoPE	2D RoPE	–	✓			IN1k	–	33.50	20.71	11.35	
+3D RoPE	3D RoPE	–	✓			IN1k	–	39.92	22.74	12.00	
+normals	3D RoPE	normals	✓			IN1k	–	38.38	12.80	5.80	**intra-patch** geometry
+depth emb.	3D RoPE	depth emb.	✓			IN1k		40.16	17.90	10.50	
+iBOT	3D RoPE	depth emb.	✓	✓		IN1k		40.45	16.52	8.40	**objective**
+MVI	3D RoPE	depth emb.	✓	✓	✓	IN1k+MVI2.0		38.58	25.29	13.19	
+dropout	**3D RoPE**	**depth emb.**	✓	✓	✓	**IN1k+MVI2.0**	✓	**40.27**	**25.25**	**16.52**	**DINOcular**
−3D RoPE	2D RoPE	depth emb.	✓	✓	✓	IN1k+MVI2.0	✓	35.54	23.25	14.58	
−depth emb.	3D RoPE	–	✓	✓	✓	IN1k+MVI2.0	–	40.92	23.15	11.73	
−DINO −iBOT	3D RoPE	depth emb.			✓	IN1k+MVI2.0		28.54	19.99	10.80	
−iBOT	3D RoPE	depth emb.	✓		✓	IN1k+MVI2.0	✓	31.92	19.87	13.60	**removal**
−iBOT −drop	3D RoPE	depth emb.	✓		✓	IN1k+MVI2.0		31.20	20.58	11.40	
−MVI	3D RoPE	depth emb.	✓	✓		IN1k+MVI2.0		34.87	15.55	4.88	
−dropout	3D RoPE	depth emb.	✓	✓	✓	IN1k+MVI2.0		38.58	25.29	13.19	

Table 2: **Ablation of the multi-view loss objective.** We finetune a baseline model that was trained with the DINO objective with different multi-view objectives. We report RGB-D semantic segmentation probing (Segm.) on NYUDepthv2 Silberman et al. (2012) (mIoU) and correspondence estimation with viewpoint change (3D Corresp.) on NAVI Jampani et al. (2023) (θ_{90}^{120}) following Probe3D El Banani et al. (2024).

Finetuning Objective	Segm. [mIoU]	3D Corresp. [R]
Baseline (trained w/ DINO objective)	40.16	17.90
MV-Cosine	11.10	11.41
DINO + MV-Cosine	24.40	7.45
MV-iBot	30.30	15.85
DINO + MV-iBot	34.00	16.86
MV-Contrastive	28.54	19.99
DINO + MV-Contrastive	31.20	20.58

results in our self-supervised setting. However, we find the simple lightweight embedding of the depth values to be reasonably effective, most notable from the depth reconstruction. It does not surpass the setting without any intra-patch depth, but we address this aspect in the design of the full training objective.

4.4 Ablation: Combination of Loss Objectives

In Table 2, we investigate the different options for multi-view consistency losses, as introduced in Section 3.2. For all objectives, we observe a collapse in segmentation performance when finetuning purely on the multi-view objective. This is expected to a certain degree, as the objective optimizes a different goal. The effect can also, as the results show, be compensated by training on a combined objective. However, what is unexpected is the drop in 3D correspondence estimation for MV-Cosine and MV-iBOT. On closer inspection, we find that Koch et al. (2025) use this loss in ensemble with 3 distillation losses that likely regularize the features well and therefore can prevent such collapse. In our setting where we do not distill from a frozen teacher, the objective is not effective. The contrastive objective is also used in (You et al., 2025) as an isolated objective and also in our case is the only one that achieves an improvement in 3D correspondence estimation. It also shows to be compatible with the DINO objective, as a combination of both not only compensates the loss in segmentation performance, but further enhances the 3D correspondence. Based on the ablation insights, we further investigate the combination of multi-view contrastive loss with semantic loss in Table 1.

Table 3: **RGB-D semantic segmentation** results (mIOU) using linear probing on the respective features. All results are reported in [% mIoU]. † marks models trained with supervised objectives, all other methods are self-supervised. • marks models with RGB and depth as input, ○ only with RGB. ⋆Sonata is not intended to be tested on these evaluation protocols.

Method	Param.	Data	ADE20k	NYUDepthv2	SUN RGB-D	Cityscapes
SoTA ViT-B / PTv3 models						
○ DINOv2 (ViT-B)	86 M	142 M	47.30	56.04	52.26	69.40
○ DINOv3 (ViT-B)	86 M	1689 M	51.80	60.78	54.28	71.53
○ DUNE (ViT-B)	85 M	21 M (dist.)	44.90	68.20	50.22	70.60
• Sonata⋆ (PTv3)	108 M	140 k (pcd)	11.53	25.67	26.76	-
ViT-B / RMT-L models at comparable data size						
○ DINO (ViT-B)	85 M	1.3 M	31.80	34.49	37.13	56.90
• MultiMAE (ViT-B)	85 M	1.3 M	17.37	30.59	30.67	28.63
• DFormerv2-L†	95.5 M	1.3 M	33.84	37.72	37.66	60.64
○ DINOv2 (ViT-B repr., only ImgNet)	85 M	1.3 M	29.83	36.16	35.90	52.14
• DINOcular-L *(Ours)*	94 M	3.9 M	**40.23**	**45.87**	**42.54**	**62.32**
ViT-S / RMT-S models at comparable data size						
○ DINO (ViT-S)	21 M	1.3 M	22.82	35.97	35.35	45.84
• DFormerv2-S†	27 M	1.3 M	30.60	23.04	28.90	51.41
○ DINOv2 (ViT-S repr., only ImgNet)	21 M	1.3 M	24.22	33.43	33.91	47.03
○ DINOv2 (ViT-S repr.)	21 M	3.9 M	20.64	29.74	31.67	44.14
○ Ours w/ RGB-only	27 M	3.9 M	18.47	28.15	28.98	39.85
• Ours w/o multi-view	27 M	1.3 M	**36.02**	**41.47**	38.35	**56.31**
• DINOcular-S *(Ours)*	27 M	3.9 M	33.88	40.82	**39.77**	55.35

Table 4: **Geometric Tasks:** We evaluate features through the Probe3D protocol on 3D Correspondence Estimation (NAVI and ScanNet) over different angles between viewpoints, as well as through one-shot Object Pose Estimation. To analyze correlation of feature information with the depth input, we report results on linear probing for depth regression/reconstruction (NYU Depth v2). † marks models trained with supervised objectives, all other methods are self-supervised. • marks models with RGB and depth as input, ○ only with RGB. ⋆Sonata is not intended to be tested on these evaluation protocols.

	NAVI [% R]				ScanNet [% R]				OnePose-LowTex [% R]			NYUDv2
Model	θ_0^{30}	θ_{30}^{60}	θ_{60}^{90}	θ_{90}^{120}	θ_0^{15}	θ_{15}^{30}	θ_{30}^{60}	θ_{60}^{180}	1 cm/1 °	3 cm/3 °	5 cm/5 °	[m RMSE] ↓
SoTA ViT-B / PTv3 models												
○ DINOv2 (ViT-B)	92.3	67.9	45.4	32.0	37.0	27.5	19.7	11.2	10	51	70	.39
○ DINOv3 (ViT-B)	96.3	71.8	41.9	25.4	54.9	39.2	26.6	15.7	4	32	54	.38
○ DUNE (ViT-B)	94.8	60.4	29.4	17.2	34.6	25.9	17.9	7.6	10	50	66	.36
• Sonata⋆ (PTv3)	73.9	41.4	27.5	23.8	37.3	24.0	11.4	4.2	0	3	8	-
ViT-B / RMT-L models at comparable data size												
○ DINO (ViT-B)	89.0	54.0	30.7	21.0	45.0	34.4	22.6	10.7	6	34	50	.64
• MultiMAE (ViT-B)	85.1	39.3	18.7	11.1	7.5	6.9	6.1	2.5	3	7	18	.19
• DFormerv2-L†	74.7	42.9	26.0	19.4	47.0	38.0	24.4	11.0	8	32	43	.72
○ DINOv2 (ViT-B repr., only ImgNet)	**93.1**	**57.4**	**31.2**	19.5	51.5	40.2	24.8	9.8	7	40	57	.63
• DINOcular-L *(Ours)*	88.2	52.7	30.3	**23.2**	**62.7**	**52.2**	**32.7**	**13.1**	**10**	**46**	**63**	.26
ViT-S / RMT-S models at comparable data size												
○ DINO (ViT-S)	86.7	52.9	31.9	21.3	41.5	30.9	19.0	9.1	6	32	47	.70
• DFormerv2-S†	80.0	41.4	24.0	17.4	37.2	26.9	16.8	7.3	**9**	32	44	.70
○ DINOv2 (ViT-S repr., only ImgNet)	88.1	54.3	29.4	19.9	50.8	39.8	24.5	9.9	7	**40**	**56**	.70
○ DINOv2 (ViT-S repr.)	84.3	52.3	29.5	17.2	51.4	40.4	25.7	11.7	6	30	46	.73
○ Ours w/ RGB-only	81.2	**58.5**	**37.0**	23.4	47.0	33.5	22.5	11.9	6	33	48	.72
• Ours w/o multi-view	87.5	49.3	25.8	16.5	44.0	35.4	22.0	8.4	7	35	51	.26
• DINOcular-S *(Ours)*	**89.1**	57.3	33.9	**25.3**	**58.3**	**46.4**	**30.3**	**16.5**	6	37	55	.24

4.5 Semantic Segmentation

An established downstream task for RGB-D perception is semantic segmentation. This task also allows us to evaluate over different data sources that are captured with different depth sensors: NYU Depth V2 (Silberman et al., 2012) and SUN RGB-D (Song et al., 2015) are based on the first generation of Kinect with structured light, Cityscapes (Cordts et al., 2016) has a stereo camera, and ADE20k has no actual depth available. We only densify sensor depth for NYU Depth v2 and SUN RGB-D, and process ADE20k with MapAnything (Keetha et al., 2026). For Cityscapes, we acquire dense depth from FoundationStereo (Wen et al., 2025), leading to three diverse depth sources. For all methods and datasets, we fit a linear probe over their training data, keeping the feature predictor completely frozen. The results are reported in Table 3.

Our features show significantly better probing results compared to both Dformerv2 and DINOv2. In most cases, DINOcular_S even outperforms the 3 times larger DINO ViT-B, while our larger model keeps advantage over models with comparable data and parameter size. However, the reference results of other large representation learners show a clear advantage of data scaling. Our results

highlight the value of geometric priors for semantic understanding. Our proposed inter-patch and intra-patch depth fusion alone yields clear gains over DINOv2 and DFormerv2 at the same data scale, and these gains persist in our full model. The multi-view objective mainly strengthens the 3D-awareness of the representations, with only slight shifts on most semantic probes. To verify that the performance gains do not result from the choice of depth source, we report additional results with diverse depth inputs in Appendix A to show robustness of our model.

4.6 3D Correspondence Estimation

To evaluate the consistency of features over different observations of the same object, we test the matching of features between pairs of images against their true correspondence based on the known object geometry. Following the evaluation setup of El Banani et al. (2024), we test this on object-centric observations from NAVI (Jampani et al., 2023) and scene-level observations on ScanNet (Dai et al., 2017). We report results grouped by the angle change between viewpoints in Table 4. Especially for large viewpoint changes, we observe a clear benefit of the multi-view training over both datasets. Remarkably, on Probe3D-ScanNet, DINOcular-S features outperform even all reference models, including DINOv3 ViT-B. Since this stands out from the more object centric NAVI data, a possible reason is that for the cluttered and larger scale environments of ScanNet, the depth provides a more noticeable advantage.

4.7 3D Object Pose Estimation

3D Object Pose Estimation requires a model to predict, in the coordinate frame of the camera, where an object is and how it is oriented. We evaluate single-shot, CAD-model-free object pose estimation for low-texture objects with the different frozen backbones on the dataset (He et al., 2022b) and evaluation protocol of You et al. (2025). We report recall for different thresholds of estimation error. We find that, in line with the results of Section 4.6, DINOcular-L features outperform all baselines and even the DINOv3 ViT-B model on the two coarser thresholds. However, there is an apparent weakness of the smaller model on the $1\,\mathrm{cm}$-1° threshold, potentially related to the blurriness of the features over the image plane that is also apparent in Figure 1.

5 Limitations

Our study focuses on moderate data scale, and it does not yet validate whether the proposed method preserves the same gains when scaled to substantially larger and diverse data. When compared at this data scale, DINOcular outperforms RGB-D and RGB-only foundation models including the DINO model family. However, it is hard to fully compare models at the same data scale. On one hand, this is because DINOv2 and DINOv3 do not make their training data available. On the other hand, DINOcular requires additional multi-view data, which in turn our ablation shows reduces performance of DINO models, hence limiting the insight from a comparison on exactly matched training data. Finally, although our learned features transfer across multiple geometric and semantic tasks, we do not observe a two-sided benefit between spatial and semantic specialization: the multi-view objective mostly affects the 3D-awareness of the representations, while there is very little change on most semantic probes and the benefit of the added depth is already apparent without the objective.

6 Conclusion

We present DINOcular, a framework for learning visuospatial features from RGB-D observations in a self-supervised manner. We demonstrate that these features generalize across semantic and geometric tasks. Our results show a significant advantage over other approaches at the same data and model scale. This is promising and motivates future research into scaling the proposed approach to larger data sources and models. Looking further ahead, this work shows that it is possible to learn generalizing features that combine complementary information from RGB and Depth inputs. This may help increase the spatial understanding of VLMs and VLAs, eventually contributing to deploy more generalizing robots to real-world environments.

AI USE STATEMENT

We have not used generative AI tools for: help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. And the following: Generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, assist with translation are not applicable to this work.
We used generative AI tools for: support in creating figures, support in editing text, finding related work, edit software code. We have reviewed all AI-assisted work. LLM-generated figures were verified for correctness. LLM-generated findings of related work were manually verified for existence and relevance. LLM-generated code edits were manually verified, tested and executed. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

For reproducibility, we provide the source code as the supplementary material (Appendix E). Details of the data processing are presented in section 4.1 and Appendix F.

REFERENCES

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European conference on computer vision*, pp. 348–367. Springer, 2022.

Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *European conference on computer vision*, pp. 677–694. Springer, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations*, volume 2026, pp. 138846–138923, 2026.

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 International Conference on Computer Vision (ICCV)*, 2021.

Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *European conference on computer vision*, pp. 561–577. Springer, 2020.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

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. *ICLR*, 2021.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21795–21806, 2024.

Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, and Ran He. Rmt: Retentive networks meet vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5641–5651, 2024.

Zhiyuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shubin Shi, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, et al. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. *arXiv preprint arXiv:2510.19400*, 2025.

RC Feord, ME Sumner, S Pusdekar, L Kalra, PT Gonzalez-Bellido, and Trevor J Wardill. Cuttlefish use stereopsis to strike at prey. *Science advances*, 6(2):eaay6036, 2020.

Robert Fox, Stephen W Lehmkuhle, and Robert C Bush. Stereopsis in the falcon. *Science*, 197(4298): 79–81, 1977.

Yan Gong, Jianli Lu, Yongsheng Gao, Jie Zhao, Xiaojuan Zhang, and Susanto Rahardja. Diff-pixelformer: Differential pixel-aware transformer for rgb-d indoor scene segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Xiaoguang Han, Yushuang Wu, Luyue Shi, Haolin Liu, Hongjie Liao, Lingteng Qiu, Weihao Yuan, Xiaodong Gu, Zilong Dong, and Shuguang Cui. Mvimgnet2. 0: A larger-scale dataset of multi-view images. *arXiv preprint arXiv:2412.01430*, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022a.

Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022b.

Christopher P Heesy. Seeing in stereo: the ecology and evolution of primate binocular vision and stereopsis. *Evolutionary Anthropology: Issues, News, and Reviews*, 18(1):21–35, 2009.

Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE international conference on image processing (ICIP)*, pp. 1440–1444. IEEE, 2019.

Krishna Jaganathan and Patricio Vela. Geomprompt: Geometric prompt learning for rgb-d semantic segmentation under missing and degraded depth. *arXiv preprint arXiv:2604.11585*, 2026.

Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin Brualla, Kaushal Patel, et al. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. *Advances in Neural Information Processing Systems*, 36:76061–76084, 2023.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction; map-anything. github. io. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509. IEEE, 2026.

Sebastian Koch, Johanna Wald, Hidenobu Matsuki, Pedro Hermosilla, Timo Ropinski, and Federico Tombari. Unified semantic transformer for 3d scene understanding. *arXiv preprint arXiv:2512.14364*, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *Advances in Neural Information Processing Systems*, 37:76819–76847, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. *arXiv preprint arXiv:1806.03198*, 2018.

Mert Bülent Sarıyıldız, Philippe Weinzaepfel, Thomas Lucas, Pau De Jorge, Diane Larlus, and Yannis Kalantidis. Dune: Distilling a universal encoder from heterogeneous 2d and 3d teachers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30084–30094, 2025.

Connor Schenck, Isaac Reid, Mithun George Jacob, Alex Bewley, Joshua Ainslie, David Rendleman, Deepali Jain, Mohit Sharma, Avinava Dubey, Ayzaan Wahid, et al. Learning the ropes: Better 2d and 3d position encodings with string. *arXiv preprint arXiv:2502.02562*, 2025.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Georgios Tziafas and Hamidreza Kasaei. Early or late fusion matters: Efficient rgb-d fusion in vision transformers for 3d object recognition. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9558–9565. IEEE, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14549–14560, 2023.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024.

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12186–12195, 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. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5249–5260, 2025.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22193–22204, 2025a.

Ziyang Wu, Jingyuan Zhang, Druv Pai, XuDong Wang, Chandan Singh, Jianwei Yang, Jianfeng Gao, and Yi Ma. Simplifying dino via coding rate regularization. *arXiv preprint arXiv:2502.10385*, 2025b.

Bo-Wen Yin, Jiao-Long Cao, Ming-Ming Cheng, and Qibin Hou. Dformerv2: Geometry self-attention for rgbd semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19345–19355, 2025.

Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. In *International Conference on Learning Representations*, volume 2024, pp. 51803–51825, 2024.

Yang You, Yixin Li, Congyue Deng, Yue Wang, and Leonidas Guibas. Multiview equivariance improves 3d correspondence understanding with minimal feature finetuning. In *International Conference on Learning Representations*, volume 2025, pp. 58081–58094, 2025.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Forrester Cole, Deqing Sun, Ming-Hsuan Yang, et al. Monst3r: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations*, volume 2025, pp. 82863–82886, 2025.

Yujia Zhang, Xiaoyang Wu, Yixing Lao, Chengyao Wang, Zhuotao Tian, Naiyan Wang, and Hengshuang Zhao. Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. *Advances in Neural Information Processing Systems*, 38:69498–69522, 2026.

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.

A Different Depth Sources as Input

We test inference on 3 different sources of depth: 2 monocular models - MapAnything and DepthAnything3. We do not observe strong performance variations between these depth sources. Due to the nature of the sensing setups, a controlled comparison between stereo and time-of-flight sensors is not possible. However, we present an additional comparison of raw sensor depth (some pixels without measurement) vs densified sensor depth vs monocular depth on NYU in Table 5.

B Efficiency

We measure performance of three models that share core RGB backbone architecture in Table 6. Two baselines are: RGB-only follows Fan et al. (2024) and DFormerv2 (Yin et al., 2025) with added depth-based biasing to the Manhattan self-attention proposed in RGB-only backbone. DINOcular, with intra-patch and inter-patch geometric priors, is faster and uses less memory than DFormerV2, with a modest latency and memory overhead relative to the RGB-only backbone.

Table 5: We probe our models on RGB-D Semantic Segmentation for ADE20k and NYU Depth v2, where depth needs to be predicted by monocular models or provided by sensors. We ablate MapAnything (Keetha et al., 2026) with DepthAnything3 (Lin et al., 2025) and nearest-neighbor densified sensor depth. The probing of the resulting features appears to be robust to variations in the depth input.

Evaluation	Depth input	DINOcular [mIoU]
ADE20k	MapAnything	33.88
ADE20k	DepthAnything3	32.68
NYU Depth v2	MapAnything	40.27
NYU Depth v2	Sensor depth, nearest-neighbor densified	39.01
NYU Depth v2	Sensor depth, densified w/ MapAnything	40.82

Table 6: Performance and efficiency comparison. Latency and memory are reported with relative differences compared to Dinocular.

Model	Parameters (M)	Average Latency (ms)	Peak Memory (MB)
Dinocular	26.7	51.07	249.12
DFormerV2	26.7	53.14 (+4.05%)	289.35 (+16.15%)
RGB-only	26.7	47.14 (-7.70%)	245.73 (-1.36%)

C Baselines Details

We compare against two types of methods: those learning features that combine spatial and visual-semantic information, and self-supervised visual representation learning.

- **Dformerv2** (Yin et al., 2025) proposes an architecture that combines RGB and Depth input to learn a general representation. They train fully supervised on image classification. This method is most directly comparable as we adapt a similar RGB architecture to the RGB-D setting and train the same model size on a similar scale of data.
- **Sonata** (Wu et al., 2025a) is a self-supervised method to learn features for point clouds. It is the only prior work of a self-supervised method for visuospatial features. However, it is meant to be applied on merged scene-level pointclouds with more even density. Comparing it on the single-view data that our model digests is ill posed and we merely report its performance as reference.
- **DINO family** (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025) is a family of models trained in self-supervised fashion on RGB only. We directly adopt their loss. DINO trains on ImageNet-1k, same as Dformer and our method and therefore is well comparable. However, for DINOv2 and DINOv3, from where we adopt the full loss, their datasets are $\times 36$ and $\times 433$ larger and not made public. Therefore, we try to reproduce results of DINOv2 in our settings or test public checkpoints with comparable data scale (Wu et al., 2025b).
- **DUNE** (Sarıyıldız et al., 2025) also only takes RGB as input, but distills a feature from both DINOv2 and Mast3r, therefore aiming to add more geometric information to its features. It is distilled from DINOv2 on an additional set of $21\,\mathrm{M}$ images, so we report its results only for reference.

D Qualitative Results

We present a qualitative evaluation of DINOcular across multiple downstream tasks, illustrating its learned feature representations, 3D geometric awareness, and 2D semantic segmentation capabilities in diverse environments.

3D Correspondence and Feature Consistency. To evaluate the model's geometric awareness and cross-view consistency, we visualize 3D correspondence matching on the NAVI dataset following the Probe3D protocol (Figure 4). DINOcular demonstrates exceptional robustness to severe viewpoint

variations, establishing dense and highly accurate matches, where baseline models such as DINO ViT-B and DUNE struggle.

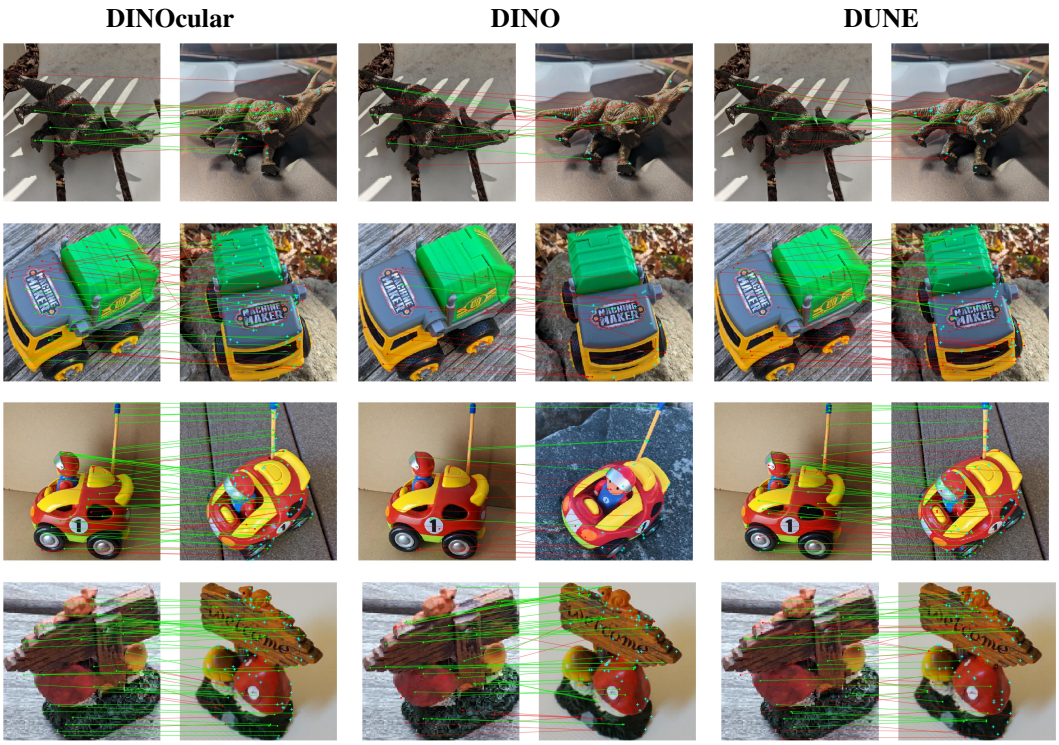

Figure 4: Qualitative comparison of 3D correspondences. DINOcular (ours) demonstrates superior robustness to large viewpoint variations, yielding more accurate matches (green) with significantly fewer outliers (red) than DINO and DUNE.

This robust 3D awareness is further reflected in the feature representations. Figure 5 visualizes the three strongest PCA components of the extracted features. DINOcular yields distinct, structurally aligned feature maps that remain highly stable across varying environments and viewpoints. It effectively achieves clear semantic part separation (e.g., distinguishing the horns, body, and legs of the object) while overcoming the noisy, pixelated artifacts prevalent in the feature spaces of both DINO and DUNE.

Semantic Segmentation. We extend our qualitative analysis to 2D semantic segmentation on complex indoor scenes shown in Figure 6. When compared to DFormerv2 (Yin et al., 2025) and MultiMAE (Bachmann et al., 2022), which exhibit severe spatial fragmentation and noisy patch assignments (particularly on large textured surfaces like floors and furniture), DINOcular generates clean, contiguous semantic regions with reliably delineated sharp object boundaries.

Finally, we visualize DINOcular's semantic segmentation performance on complex outdoor driving scenes in Figure 7. Model demonstrates successful generalization for outdoor urban layouts, correctly identifying primary categories. Notably, DINOcular exhibits strong robustness to challenging, real-world illumination shifts; it maintains coherent structural parsing even in the presence of lighting changes, e.g. shadows.

E CODE

We provide our code in an anonymized version as a supplement to this submission.

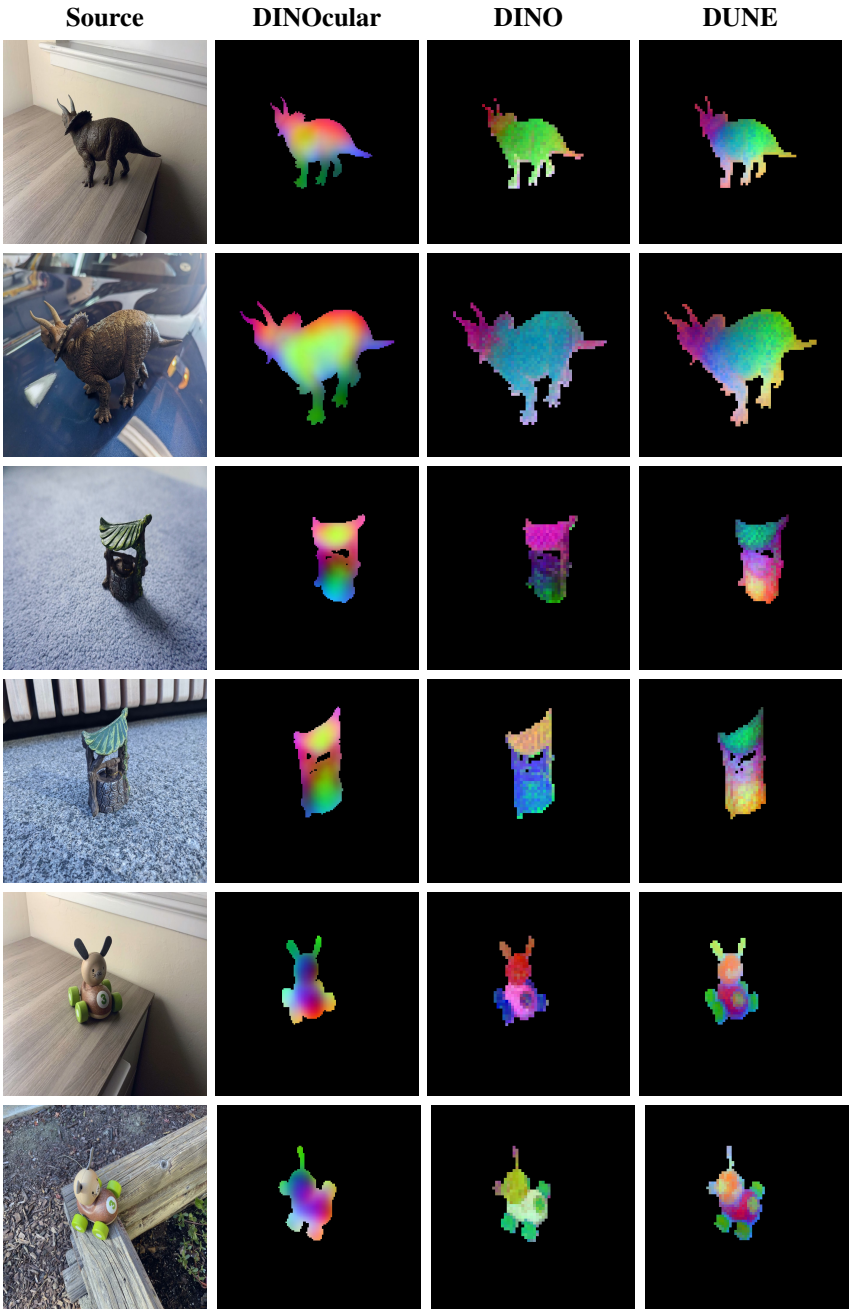

Figure 5: PCA visualization of extracted features. DINOcular demonstrates smoother, cross-view consistent features and clear semantic part separation.

F Training Details

By default we train all models on 4 nodes with each 4 A100 GPUs. The details are shown in Table 8. Augmentation and local-global crop sampling follows DINOv2 (Oquab et al., 2023) regime for the RGB input. The depth input passed as raw metric values without any normalization and simply matches the final augmented RGB input size and rotation. MVImgnet2.0 (Han et al., 2024) is an object-centric dataset, with a large portion of redundant background information, which is not useful for multi-view correspondence loss. To constrain the objective's focus on the object we predict object masks using SAM3 (Carion et al., 2026). We do not observe any semantic prior benefits from

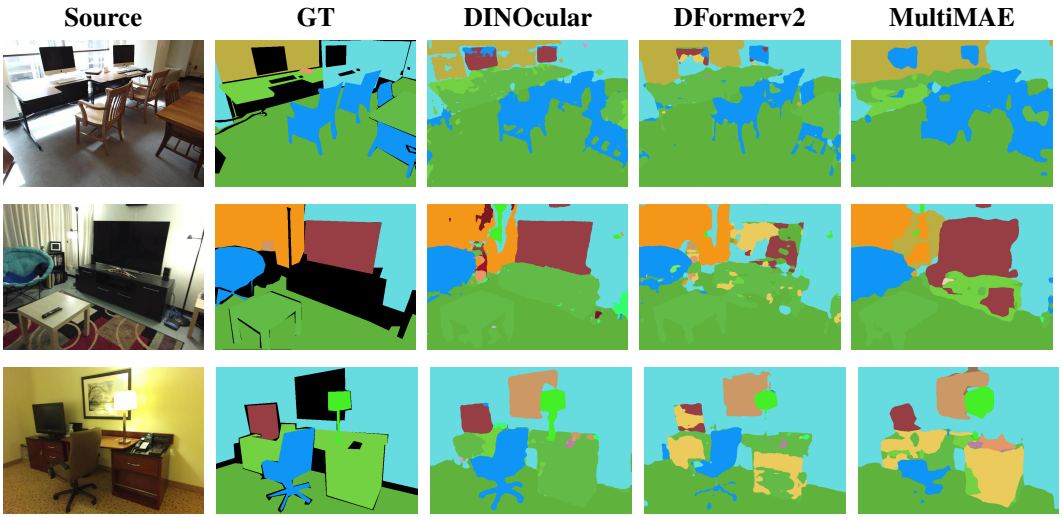

Figure 6: Qualitative semantic segmentation comparison. DINOcular yields highly consistent semantic masks and sharper object boundaries compared to DFormerv2 and MultiMAE.

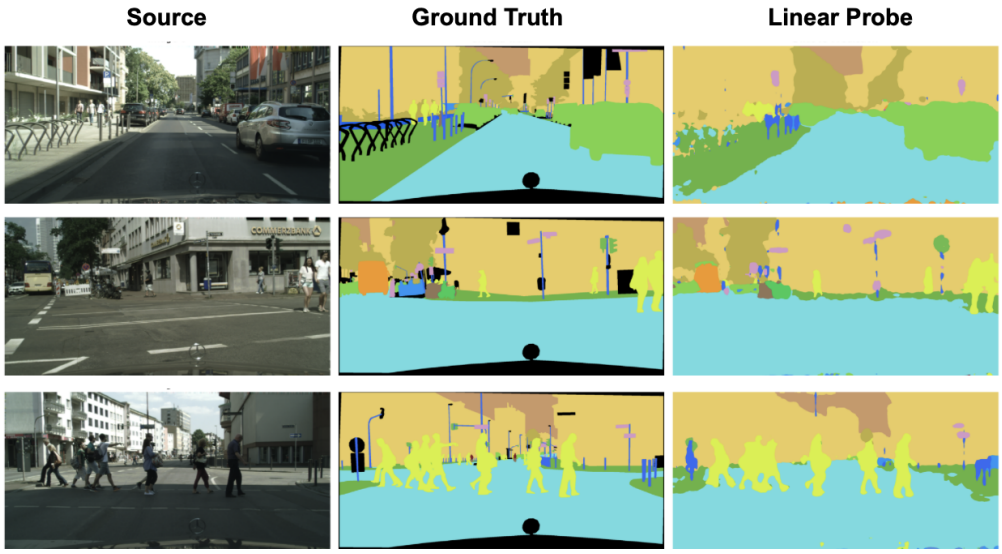

Figure 7: Qualitative semantic segmentation results on CityScapes outdoor scenes for DINOcular. The model reliably captures spatial layout and remains robust to lighting changes, e.g. shadows.

SAM3 (Carion et al., 2026), and it merely helps to focus multi-view loss on the object. The ablation results are reported in Table 7.

G Experiment details

For evaluation on the downstream tasks, we probe frozen encoders with the same evaluation recipes, ensuring that DINOcular performance gains come from the proposed architecture and pretraining framework.

Table 7: Influence of using object masks during pre-training.

Masks	**Segm.** [mIoU]	**NAVI** θ_{90}^{120} [R]	**ScanNet** θ_{60}^{180} [R]
No	39.89	22.50	18.80
Yes	40.27	25.25	16.52

Table 8: Hyperparameters and training details used for pretraining. Highlighted rows indicate difference with model size.

Parameter	**Value**
Architecture & Data	
Base Architecture	DINOcular_S/_L
Projection dimension	65,536
Mixed precision	BF16
Batch size per GPU	96/46
Total epochs	150
Optimization	
Optimizer	AdamW
Peak learning rate	0.004
Learning rate scaling	$\text{lr} \times \sqrt{\text{batch_size}/1024}$
Min learning rate	1×10^{-6}
Warmup epochs	15
Weight decay	$0.04 \rightarrow 0.4$
Gradient clipping norm	3.0
Freeze last layer	1 epoch
Stochastic depth (Drop path) rate	0.1/0.2
Self-Supervised Objectives	
Teacher momentum	$0.992/0.996 \rightarrow 1.0$
Teacher temperature	$0.04 \rightarrow 0.07$
Teacher temperature warmup	15 epochs
iBOT student temperature	0.1
iBOT patch mask ratio	[0.1, 0.5]
iBOT mask probability	0.5
DINO loss weight	1.0
iBOT patch loss weight	1.0
KoLeo entropy loss weight	0.1
Multi-view loss weight	10.0
Multi-Crop Augmentation	
Global crops number	2
Global crops scale	(0.32, 1.0)
Local crops number	8
Local crops scale	(0.05, 0.32)