DINOv3D: 2D-3D Joint Optimization for Unified Spatial Understanding

Bo Zhou[1,3,7†], Jianzhe Gao[2,3†], Zhihui Wang[3], Lingxiang Wu[3,4,6], Jinqiao Wang[3,4,5,6], Yazhou Yao[1,7], and Wenguan Wang[2✉]

[1] Nanjing University of Science and Technology, Nanjing, China
[2] The State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China
[3] Zidong Taichu Technology Co., Ltd., Beijing, China [4] Wuhan AI Research, Wuhan, China
[5] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[6] Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[7] State Key Laboratory of Intelligent Manufacturing of Advanced Construction Machinery, Nanjing, China
https://github.com/bobochow/DINOv3D

Abstract. Adapting 2D foundation models for 3D spatial understanding faces a critical dilemma. Fine-tuning 2D models to learn 3D geometry causes the catastrophic forgetting of native 2D knowledge. Conversely, freezing the 2D model restricts 3D spatial perception. To resolve this issue, we propose DINOv3D, a joint optimization framework built upon DINOv3 that employs a homologous teacher-student architecture to establish a regularized integration between the 2D and 3D understanding. To preserve original 2D priors, DINOv3D distills 2D knowledge into a 3D Gaussian regularization field, which aligns the rendered features with the reference visual features from a frozen teacher model. Meanwhile, a student model processes long-context multi-view inputs through parameter-efficient fine-tuning. This step injects 3D spatial consistency priors to 2D model while using the regularization field to mitigate the forgetting of 2D knowledge. To enrich the hierarchical representation of 3D spatial understanding, the student model predicts additional semantic and instance Gaussian features. We then apply a ray-depth-semantic alignment mechanism, which uses 3D depth-ray priors to enforce multi-view consistency across 2D semantic rendering. Extensive experiments demonstrate that, despite updating only 10% of the parameters of a 1B-parameter foundation model VGGT, DINOv3D achieves state-of-the-art results in comprehensive 3D scene understanding across challenging novel view synthesis, depth estimation, and open-vocabulary semantic segmentation. Furthermore, DINOv3D consistently enhances the generalizability of the DINOv3 backbone on 2D linear probing benchmarks.

Keywords: Representation Learning · Foundation Models · Scene Understanding · Gaussian Splatting · Joint Optimization

1 Introduction

Modern perception systems increasingly rely on 2D Vision Foundation Models (VFMs) for diverse visual tasks [5, 6, 13, 19, 41, 45, 50, 54, 66]. To adapt 2D VFMs

† Work done during internship at Zidong Taichu.
✉ denotes corresponding authors.

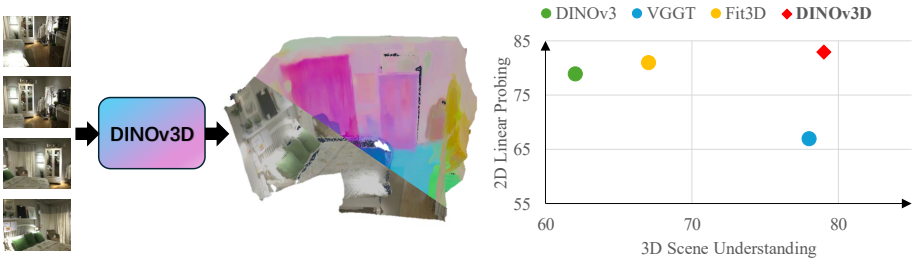

Fig. 1: DINOv3D takes multi-view images as input and produces a unified 3D Gaussian scene representation, achieving state-of-the-art results across 2D and 3D domains.

to the 3D domain, conventional practices involve extracting features using frozen 2D models augmented with additional 3D modules [71], or performing unconstrained fine-tuning combined with 3D Gaussian Splatting [7, 9, 11, 16, 29, 42, 53, 55, 73]. However, as these 2D models are trained on unposed images, they inherently lack multi-view consistency and 3D spatial awareness [23, 47, 57, 68]. Consequently, this paradigm often exhibits inconsistencies under viewpoint variations and occlusions [18, 37, 47, 57]. To overcome this limitation, the adaptation requires the integration of 3D geometric regularization, which enforces spatial consistency and provides the structural context for scene understanding [14, 23, 47, 55, 57, 68].

Realizing this geometry-regularized adaptation presents significant optimization challenges. The naive fine-tuning of 2D models causes the catastrophic forgetting of inherent 2D knowledge. While standard teacher-student distillation from 2D models mitigates this forgetting, it relies on 3D geometric pseudo-labels [68]. Furthermore, this distillation exposes an inherent defect. Severe occlusions within short-context inputs lead to unreliable 3D geometric constraints, thereby injecting erroneous distillation gradients into the student model. Consequently, the regularized adaptation necessitates the capability to process long-context inputs. This capability provides spatial-temporal geometric consistency, which acts as a reliable physical filter to correct noisy 2D priors and ensures that the student model learns robust 3D structures without collapsing [57].

To address these challenges, we introduce DINOv3D, a unified spatial understanding framework based on DINOv3 [50]. Built upon a homologous teacher-student dual-branch architecture, DINOv3D leverages 3D Gaussian Splatting [29] as a core representation to enable both 2D prior-anchored regularization and long-context 3D prior-guided multi-view consistent alignment. Specifically, a frozen teacher model distills inherent 2D knowledge into a 3D Gaussian regularization field by enforcing rendering consistency against the 2D feature maps of the teacher. This field serves as a reliable anchor to prevent catastrophic forgetting. Concurrently, the corresponding student model is updated via LoRA [24] to absorb new 3D geometric constraints. Unlike the frozen teacher, which processes individual frames independently to preserve pure 2D spatial priors, the student model establishes multi-view consistent alignment by incorporating a shifted window attention mechanism during feature extraction to build cross-frame connections. Furthermore, DINOv3D utilizes 3D Gaussian Splatting to

represent and render the extracted semantic features within this unified regularization field. To prevent semantic inconsistency during the rendering process, a ray-depth-semantic alignment strategy utilizes depth-ray 3D priors to explicitly constrain the cross-view consistency of the regularization field.

Extensive experiments demonstrate the superiority of DINOv3D across 3D and 2D domains. For 3D scene understanding, DINOv3D achieves state-of-the-art performance in novel view synthesis, depth estimation, and open-vocabulary semantic segmentation on ScanNet [12]. Specifically, in novel view synthesis, DINOv3D outperforms Uni3R [53] by $+$**1.78** dB PSNR, using merely **10**% trainable parameters. Furthermore, DINOv3D preserves 2D knowledge across diverse linear probing tasks, including global classification and dense prediction. For instance, it yields a $+$**1.3**% improvement over DINOv3 [50] on ImageNet-R [21].

2 Related Work

3D Scene Understanding. The paradigm of 3D scene understanding is actively transitioning from computationally expensive per-scene optimization [2, 8, 10, 22, 29, 67, 73] to efficient feed-forward frameworks [4, 7, 27, 34, 38, 58–60, 70]. Although recent unified models [15, 48, 53, 55, 56] jointly formulate geometric and semantic fields from sparse views, inherent memory constraints restrict the operational capacity to short context windows (*e.g.*, 2–16 views). This restriction severely diminishes the extended multi-view spatial coverage and global geometric coherence of the reconstructed scenes, particularly under severe occlusion or within texture-less regions. Furthermore, the decoupled optimization of 2D visual representations and 3D structural priors in these conventional methods frequently induces semantic inconsistencies across varying viewpoints. The proposed memory-efficient architecture directly addresses these limitations by facilitating long-context joint optimization across the spatiotemporal domains.

Lifting 2D Features to 3D. Lifting 2D features to 3D requires transferring strong priors while preserving cross-view consistency. Early optimization-based methods [17,30,41,42,52,73,75] match 3D fields to 2D visual foundation models, but their computational cost hinders real-time application. Recent feed-forward approaches [15,33,53,56] directly unproject 2D features using explicit camera parameters. However, these deterministic projection paradigms implicitly assume that VFM features (*e.g.*, CLIP [45], DINOv2 [41], SAM [46]) are strictly consistent across viewpoints. In reality, VFMs exhibit significant view-dependent semantic instability. Existing strategies relying on simple aggregation heuristics inherently fail to dynamically resolve these semantic conflicts, inevitably causing blurred 3D representations. Our work addresses this not by altering native features before projection, but by leveraging a trainable student backbone that incorporates long-range multi-view consistency to dynamically resolve conflicts.

Injecting 3D Priors to 2D. Conversely, utilizing 3D structures to enhance 2D perception remains less explored. Early works like Pri3D [23] utilize explicit 3D data for contrastive pre-training. Recent studies transition toward implicit

supervision, demonstrating that extensive multi-view observations produce robust, view-invariant 2D features [57] and unify 2D-3D tasks [47]. However, these approaches frequently lack explicit geometric reconstruction. While FiT3D [68] incorporates differentiable rendering, it still depends on slow per-scene optimization. Furthermore, geometry-focused models like Depth Anything 3 [34] provide strong reconstruction but often compromise semantic capabilities. Our method synthesizes these paradigms by utilizing extensive multi-view sequences within a feed-forward framework to inject 3D structural cues, thereby simultaneously improving the geometric awareness and semantic generalization of 2D features.

3 Method

3.1 Problem Formulation

Given a sequence of input images $\mathcal{I}=\{\boldsymbol{I}_k \in \mathbb{R}^{H\times W\times 3}\}_{k=1}^{K}$, our primary objective is to jointly recover a 3D Gaussian representation of the scene and refine the 2D representations of the foundation model. This formulation bridges the visual space of foundation models with the geometric space of 3D Gaussians.

Visual Space. For each image $\boldsymbol{I}$, we first patch it into visual tokens $\boldsymbol{T}\in\mathbb{R}^{N\times C}$ and subsequently prepend a camera embedding token $\boldsymbol{e}_{cam}\in\mathbb{R}^9$ to visual tokens. Following Depth Anything 3 [34], a unified Dense Prediction Transformer (DPT) decoder processes these features to predict a depth map $\boldsymbol{D} \in \mathbb{R}^{H\times W}$, a dense ray map $\boldsymbol{R} \in \mathbb{R}^{H\times W\times 6}$ and a Gaussian field $\mathcal{G}$. We directly utilize the pose extraction algorithm from [34] to obtain the camera poses required for the 3D Gaussian splatting [29]. The ray map implicitly encodes camera geometry via ray origins $\boldsymbol{o_u} \in \mathbb{R}^3$ and directions $\boldsymbol{d_u} \in \mathbb{R}^3$ for each 2D pixel $\boldsymbol{u}$. The 3D coordinates are unprojected using the predicted depth and the ray directions.

Regularized 3D Gaussian Field. DINOv3D predicts a 3D Gaussian field $\mathcal{G}=\{\boldsymbol{g}\}$ in a feed-forward manner. Each Gaussian $\boldsymbol{g}$ is centered at 3D coordinate $\boldsymbol{x}\in\mathbb{R}^3$ with a rotation quaternion $\boldsymbol{q}\in\mathbb{R}^4$, opacity $\boldsymbol{\alpha}\in\mathbb{R}^+$, scales $\boldsymbol{s}\in\mathbb{R}^3$ and SH coefficients $\boldsymbol{\beta}\in\mathbb{R}^{48}$ of degree 3. To enable unified understanding and joint optimization, we attach three feature vectors to each Gaussian primitive: $\boldsymbol{f}_{sem}\in\mathbb{R}^{64}$ for open-vocabulary semantics, $\boldsymbol{f}_{ins}\in\mathbb{R}^8$ for instance discrimination, and a regularization feature $\boldsymbol{f}_{reg}\in\mathbb{R}^{64}$. Because explicitly defining semantics and instance boundaries requires specialized priors beyond the generalized visual features of DINOv3, our framework operates as a unified distillation hub. Specifically, it absorbs explicit knowledge from task-specific experts, such as LSeg [32] and SAM [6], into $\boldsymbol{f}_{sem}$ and $\boldsymbol{f}_{ins}$. Concurrently, $\boldsymbol{f}_{reg}$ serves as a distillation target aligned with a frozen teacher model to ground the intrinsic features and prevent the catastrophic forgetting of 2D knowledge. Consequently, the complete parameter set of the 3D Gaussian is defined as $\boldsymbol{g}=\{\boldsymbol{x},\boldsymbol{q},\boldsymbol{\alpha},\boldsymbol{s},\boldsymbol{c},\boldsymbol{f}_{sem},\boldsymbol{f}_{ins},\boldsymbol{f}_{reg}\}$.

Teacher-Student Regularized Architecture. As illustrated in Fig. 2, the overall architecture comprises two branches, where both the teacher and student models are initialized with the same DINOv3 [50] weights. In the teacher branch, the visual backbone of the teacher model $\mathcal{E}_\psi$ is frozen, and the task-specific DPT decoder is trained alongside the projection heads. Specifically, these projection

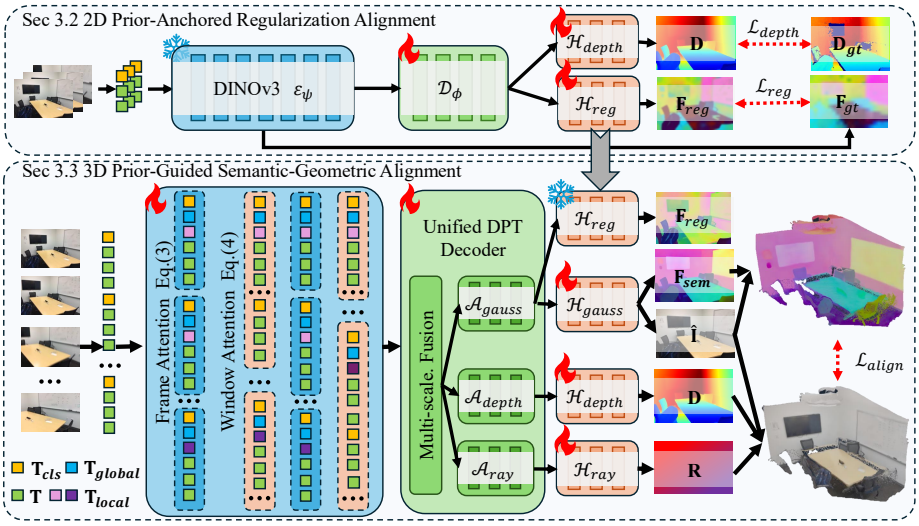

Fig. 2: Overview of DINOv3D framework. DINOv3D lifts 2D DINOv3 features into a regularized 3D Gaussian field, where a frozen teacher preserves 2D priors and a student learns semantic-geometric consistency through rendering-based alignment.

heads employ differentiable rendering to project the 3D Gaussian field into task-specific 2D rendering outputs. Subsequently, the backbone of the student is unlocked via LoRA [24] to jointly optimize the consistency of the rendering results, encompassing appearance, semantics, and regularization features. During this second branch, the regularization head $\mathcal{H}_{reg}$ is strictly frozen to enforce an invariant projection subspace. Combined with explicit supervision from the frozen teacher, this mechanism ensures that the evolving representation of the student remains firmly anchored to the original 2D knowledge.

3.2 2D Prior-Anchored Regularization Alignment

In the teacher branch, our goal is to establish a geometrically consistent 3D Gaussian representation that is strictly grounded in the feature space of the 2D foundation model. We freeze the backbone $\mathcal{E}_\psi$ of teacher to preserve 2D knowledge and optimize the decoder $\mathcal{D}_\phi$ along with the regularization head $\mathcal{H}_{reg}$.
Regularization Field Construction. To enforce consistency between the 2D features and the 3D Gaussian field, we utilize $\mathcal{H}_{reg}$ to predict features $\boldsymbol{f}_{reg}$, which are rendered into a 2D feature map $\boldsymbol{F}_{reg}$ via differentiable rasterization [29]:

$$\boldsymbol{F}_{reg} = \sum_{i=1}^{N_g} \boldsymbol{f}_{reg} \cdot \boldsymbol{\alpha}_i \prod_{j=1}^{i-1} (1 - \boldsymbol{\alpha}_j), \tag{1}$$

where N_g denotes the number of 3D Gaussians. This rendered map bridges the geometric 3D Gaussian field and the continuous 2D feature space.
Feature-Metric Alignment Loss. We treat the output features of the frozen teacher model, denoted as $\boldsymbol{F}_{gt} = \mathcal{E}_\theta(\boldsymbol{I})$, as the pseudo-ground-truth. The alignment is driven by minimizing the discrepancy between the rendered features and

the native features of the foundation model across a set of valid pixels Ω, which are identified by filtering based on the confidence of the depth map:

$$\mathcal{L}_{reg} = \frac{1}{|\Omega|} \sum\nolimits_{\boldsymbol{u} \in \Omega} \big(1 - \frac{\boldsymbol{F}^{\boldsymbol{u}}_{reg} \cdot \boldsymbol{F}^{\boldsymbol{u}}_{gt}}{\|\boldsymbol{F}^{\boldsymbol{u}}_{reg}\|_1 \|\boldsymbol{F}^{\boldsymbol{u}}_{gt}\|_1}\big). \tag{2}$$

By optimizing $\mathcal{L}_{reg}$, we constrain the decoder to align 3D Gaussian projections with high-quality 2D features, establishing a robust 2D knowledge anchor.

3.3 3D Prior-Guided Semantic-Geometric Alignment

Combined with the 2D knowledge anchor, we employ LoRA fine-tuning to capture 3D priors (*e.g.*, spatial-temporal consistency and object permanence) while simultaneously refining the 3D geometry estimation. To process long-context sequences (*i.e.*, up to 128 frames) within this phase, we implement a shifted-window attention mechanism to facilitate spatial-temporal resonance, a unified decoding architecture, and a ray-depth-semantic alignment strategy.

Spatial-temporal Resonance via Shifted Windows. Directly applying global attention across extensive input sequences by concatenating all frames is computationally prohibitive and semantically redundant. To address this limitation, we introduce a *Shifted Window Attention with Dual Memory* mechanism. This mechanism captures the hierarchical multi-view connection through the design of two distinct categories of learnable memory tokens. Specifically, the global memory token $\boldsymbol{T}_{global} \in \mathbb{R}^C$ models the scene-level representation, while local memory tokens $\boldsymbol{T}_{local} \in \mathbb{R}^{W \times C}$ provide window-level conditioning. Besides, the inherent class token $\boldsymbol{T}_{cls} \in \mathbb{R}^{K \times C}$ is preserved to extract the frame-level summary. Structurally, we partition the transformer into two sequential groups of sizes L_s and L_g. The initial L_s layers execute intra-frame self-attention independently for each image. For the k-th frame residing in a context window w, the spatial visual tokens $\boldsymbol{T}_k$, the class token $\boldsymbol{T}_{cls,k}$, and the dual memory tokens are concatenated along the sequence dimension. The feature extraction at layer $l \in \{1, \ldots, L_s\}$ utilizes the standard multi-head self-attention:

$$[\boldsymbol{T}^l_{cls,k}; \boldsymbol{T}^l_{global}; \boldsymbol{T}^l_{local,w}; \boldsymbol{T}^l_k] = \mathrm{Attn}\big([\boldsymbol{T}^{l-1}_{cls,k}; \boldsymbol{T}^{l-1}_{global}; \boldsymbol{T}^{l-1}_{local,w}; \boldsymbol{T}^{l-1}_k]\big), \tag{3}$$

where $[\cdot;\cdot]$ denotes the concatenation operation. This configuration ensures that both global and local memory tokens actively participate in the frame-wise representation learning to inject hierarchical spatial-temporal awareness.

Subsequently, the L_g layers alternate between the aforementioned intra-frame attention and a local cross-view window attention. To circumvent the quadratic complexity of global cross-view interactions, we restrict the temporal attention span by partitioning the sequence into local windows of size W. For a specific window w encompassing a subset of frame indices $\boldsymbol{Z}_w$, the cross-view attention operates via window-based self-attention on the aggregated token sequence:

$$[\boldsymbol{T}^l_{cls,w}; \boldsymbol{T}^l_{global}; \boldsymbol{T}^l_{local,w}; \boldsymbol{T}^l_w] = \mathrm{Attn}\big([\boldsymbol{T}^{l-1}_{cls,w}; \boldsymbol{T}^{l-1}_{global}; \boldsymbol{T}^{l-1}_{local,w}; \boldsymbol{T}^{l-1}_w]\big), \tag{4}$$

where $\boldsymbol{T}^l_{cls,w} = \text{Concat}_{i \in \boldsymbol{Z}_w}(\boldsymbol{T}^l_{cls,i})$ and $\boldsymbol{T}^l_w = \text{Concat}_{i \in \boldsymbol{Z}_w}(\boldsymbol{T}^l_i)$ represent the aggregation of the frame-level and spatial tokens within the window.

To establish long-range spatial-temporal dependencies without invoking expensive global attention [20, 39, 61–63], we shift the partition by $\lfloor W/2 \rfloor$ views in successive cross-view layers. Through this alternating routing scheme, $\boldsymbol{T}_{global}$ functions as a continuous global bridge, propagating global semantic information across distinct subsets of views. Concurrently, $\boldsymbol{T}_{local}$ specializes in capturing fine-grained local variations within the boundaries of a specific window. This proposed structure efficiently facilitates the joint optimization of 2D features and 3D geometry while strictly maintaining linear computational complexity.

Unified Geometric-Semantic Decoding. To facilitate the simultaneous refinement of geometric cues (*i.e.*, depth and rays) and semantic properties (*i.e.*, 3D Gaussian fields), we employ a unified DPT decoder $\mathcal{D}\phi$ as a centralized processing module. Rather than utilizing separate branches that can introduce feature misalignment, this decoder aggregates the backbone features from four scales into a single high-resolution representation $\boldsymbol{F}_{shared} = \Psi(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_4)$.

From this shared representation, lightweight convolutional refinement modules $\mathcal{A}$ first project the features into specific functional subspaces, which are then mapped by prediction heads $\mathcal{H}$ to yield dense geometric predictions:

$$\begin{gathered} \boldsymbol{D} = \mathcal{H}_{depth}(\mathcal{A}_{depth}(\boldsymbol{F}_{shared})), \quad \boldsymbol{R} = \mathcal{H}_{ray}(\mathcal{A}_{ray}(\boldsymbol{F}_{shared})), \\ \boldsymbol{G} = \mathcal{H}_{gauss}(\mathcal{A}_{gauss}(\boldsymbol{F}_{shared})). \end{gathered} \tag{5}$$

This shared architecture learns task-agnostic representations in the fused space, subsequently refined through targeted optimization in individual branches.

Ray-Depth-Semantic Alignment. Existing methods frequently lift 2D features into 3D without imposing geometric constraints, despite the importance of geometry for comprehensive scene understanding. To facilitate structurally-aware alignment, we initially address the limitations of current multi-view datasets (*e.g.*, noisy 2D labels projected from sparse 3D annotations) by constructing a high-quality, large-scale multi-view dataset, Uni3D. The data are collected from existing datasets (ScanNet [12], ScanNet++ [65], RealEstate10K [74], DL3DV [35], *etc.*). We employ a multi-stage pipeline utilizing foundation models (*e.g.*, Depth Anything 3 [34] and SAM 3 [6]) to generate dense and spatially consistent annotations for geometry, semantics, and instances. Specifically, we extract 2D pseudo-labels using these models, establish multi-view consistent 2D correspondences based on camera parameters, and generate aggregated 3D annotations via the back-projection of the 2D labels. This dataset supplies the supervision required by the alignment module (refer to the Appendix for details).

The ray-depth-semantic alignment module couples semantic learning with geometric constraints. By leveraging the predicted rays $\boldsymbol{R}$ and depth $\boldsymbol{D}$, we establish the geometric correspondence between a pixel $\boldsymbol{u}_s$ in the source view and its corresponding pixel $\boldsymbol{u}_t$ in the target view:

$$\boldsymbol{u}_{s \to t} = \boldsymbol{P}_t\big(\boldsymbol{o}_{\boldsymbol{u}_s} + \boldsymbol{D}_{\boldsymbol{u}_s} \cdot \boldsymbol{d}_{\boldsymbol{u}_s}\big), \tag{6}$$

where $\boldsymbol{P}_t$ denotes the projection operation. Since the student model is unlocked to adapt geometric knowledge, early-stage predictions are inevitably noisy. Aligning features via flawed correspondences risks injecting noise into the gradients of the model, causing feature space collapse. To ensure robust optimization, we modulate the alignment loss using the depth confidence metric $\boldsymbol{M}_{\boldsymbol{u}_s}$:

$$\mathcal{L}_{align} = \sum\nolimits_{(\boldsymbol{u}_s, \boldsymbol{u}_t) \in \mathcal{Q}} \boldsymbol{M}_{\boldsymbol{u}_s} \big(\|\boldsymbol{F}_{sem}^{\boldsymbol{u}_s} - \boldsymbol{F}_{sem}^{\boldsymbol{u}_{s\to t}}\|_1 + \|\boldsymbol{F}_{ins}^{\boldsymbol{u}_s} - \boldsymbol{F}_{ins}^{\boldsymbol{u}_{s\to t}}\|_1 \big), \tag{7}$$

where $\mathcal{Q}$ denotes the collection of pixel pairs that successfully project within the spatial boundaries of the target view. The minimization of $\mathcal{L}_{align}$ aligns semantic boundaries with geometric discontinuities, thereby enhancing the fidelity of the 3D representation and propagating geometric cues back into the 2D knowledge.

3.4 Optimization Objectives

DINOv3D is trained end-to-end using a composite loss function that jointly optimizes high-fidelity appearance, accurate geometry, semantic consistency, and instance discrimination. The total objective $\mathcal{L}_{total}$ is shown below:

$$\mathcal{L}_{total} = \mathcal{L}_{reg} + \mathcal{L}_{rgb} + \mathcal{L}_{depth} + \mathcal{L}_{ray} + \mathcal{L}_{sem} + \mathcal{L}_{ins} + \mathcal{L}_{align}. \tag{8}$$

Photometric Reconstruction ($\mathcal{L}_{rgb}$). To ensure high-quality visual synthesis, we minimize the difference between the rendered image $\hat{\boldsymbol{I}}$ and the ground truth $\boldsymbol{I}$. We employ a combination of $\mathcal{L}_1$ loss and a LPIPS loss term [69]:

$$\mathcal{L}_{rgb} = \|\hat{\boldsymbol{I}} - \boldsymbol{I}\|_1 + \mathcal{L}_{lpips}(\hat{\boldsymbol{I}}, \boldsymbol{I}). \tag{9}$$

Geometric Consistency ($\mathcal{L}_{depth}$). To supervise the explicit geometry predicted by our DPT decoder, we combine a confidence-weighted depth $\mathcal{L}_1$ term with a depth-gradient regularizer:

$$\mathcal{L}_{depth} = \frac{1}{|\Omega|} \sum\nolimits_{\boldsymbol{u} \in \Omega} \left(\boldsymbol{M}_{\boldsymbol{u}} \left| \boldsymbol{D}_{\boldsymbol{u}} - \boldsymbol{D}_{gt,\boldsymbol{u}} \right| - \log \boldsymbol{M}_{\boldsymbol{u}} \right) + \|\nabla \boldsymbol{D} - \nabla \boldsymbol{D}_{gt}\|_1. \tag{10}$$

Here, Ω represents the set of valid pixels, $\boldsymbol{M}_{\boldsymbol{u}}$ denotes depth mask, and ∇ denotes spatial finite difference operators.

Ray and Reprojection Loss ($\mathcal{L}_{ray}$). We supervise both the predicted ray map and the reprojected point map reconstructed from depth and rays:

$$\mathcal{L}_{ray} = \|\boldsymbol{R} - \boldsymbol{R}_{gt}\|_1 + \|\hat{\boldsymbol{X}} - \boldsymbol{X}_{gt}\|_1, \quad \boldsymbol{X} = \boldsymbol{o} + \boldsymbol{D} \cdot \boldsymbol{d}, \tag{11}$$

where $\boldsymbol{X}_{gt}$ is the normalized ground-truth reprojected point map, and $(\boldsymbol{o}, \boldsymbol{d})$ are the predicted ray origin and direction encoded in $\boldsymbol{R}$.

Semantic and Instance Supervision ($\mathcal{L}_{sem}$, $\mathcal{L}_{ins}$). For the semantic and instance branches, we render the feature maps $\boldsymbol{F}_{sem}$ and $\boldsymbol{F}_{ins}$ using the same Gaussian rasterizer. We supervise the semantic branch by distilling from a frozen

LSeg encoder: given LSeg features $\boldsymbol{F}_{lseg}$ extracted from the target view, we align them with the rendered semantic features $\boldsymbol{F}_{sem}$ using a cosine-similarity loss:

$$\mathcal{L}_{sem} = \frac{1}{|\Omega|} \sum\nolimits_{\boldsymbol{u}\in\Omega} \big(1 - \frac{\boldsymbol{F}^{\boldsymbol{u}}_{lseg} \cdot \boldsymbol{F}^{\boldsymbol{u}}_{sem}}{\|\boldsymbol{F}^{\boldsymbol{u}}_{lseg}\|_1 \, \|\boldsymbol{F}^{\boldsymbol{u}}_{sem}\|_1}\big), \tag{12}$$

which lifts rich 2D open-vocabulary semantics into the 3D Gaussian representation. The instance branch employs SAM [6] masks $\mathcal{S}$ for weak supervision, applying a cosine contrastive loss to the rendered feature map $\boldsymbol{F}_{ins}$ to align pixels $\boldsymbol{u}$ and $\boldsymbol{u}'$ within the same mask $\mathcal{S}$:

$$\mathcal{L}_{ins} = -\log \frac{1}{|\Omega|} \sum_{\boldsymbol{u}\in\Omega} \Big(\sum_{\boldsymbol{u}'\in\mathcal{S}} \mathrm{sim}\big(\boldsymbol{F}^{\boldsymbol{u}}_{ins}, \boldsymbol{F}^{\boldsymbol{u}'}_{ins}\big) + \sum_{\boldsymbol{u}'\notin\mathcal{S}} \big(1 - \mathrm{sim}\big(\boldsymbol{F}^{\boldsymbol{u}}_{ins}, \boldsymbol{F}^{\boldsymbol{u}'}_{ins}\big)\big)\Big). \tag{13}$$

4 Experiments

4.1 Experimental Setup

Unified 3D Scene Understanding and Visual Geometry Estimation. Following the protocol established in [15], we evaluate the model on ScanNet [12] across three tasks: novel view synthesis, depth estimation, and open-vocabulary semantic segmentation. Performance is quantified using PSNR, SSIM, and LPIPS for rendering; Absolute Relative error (Abs Rel) and Inlier Ratio ($\delta < 1.25$) for geometry; and mIoU and mAcc for semantics. Furthermore, geometric robustness is evaluated across distinct tasks and datasets. Specifically, we assess novel view synthesis on RealEstate10K [74] under various overlap conditions, relative pose estimation on RealEstate10K and ACID [36], and cross-dataset generalization on ACID and DTU [26]. For the recovery of relative poses, we report the Area Under the Curve (AUC) at thresholds of 5°, 10°, and 20°.
2D Feature Linear Probing. To verify that 3D-aware training retroactively enhances the 2D backbone, we train linear heads on the frozen features following the protocol of DINOv3 [50]. We evaluate the representation generalizability across global classification benchmarks (ImageNet-Real [3], ImageNet-R [21], ObjectNet [1], and Oxford-H [44]) and dense prediction tasks, including semantic segmentation (ADE20K [72]), depth estimation (NYUv2 [49]), video segmentation (DAVIS [43]), and geometric correspondence (NAVI [25] and SPair [40]).

4.2 Implementation Details

We implement DINOv3D in PyTorch using 8 NVIDIA H20 GPUs. The backbone network is initialized with the pre-trained DINOv3 ViT-L [50].
Training Configuration. DINOv3D is trained using the AdamW optimizer with a learning rate of 1×10^{-4}, a weight decay of 0.05, and a cosine annealing schedule. Input images are processed at a base resolution of 256×256. During the training process, the student backbone is adapted via LoRA with a rank of $r = 16$, alongside the joint training of the DPT decoder and the task-specific

Table 1: Quantitative comparison (§4.3) on ScanNet across different number of input views for novel view synthesis (NVS), open-vocabulary semantic segmentation (OVSS), and depth estimation (DE).

Methods	Params / Train Params	NVS		OVSS		DE	
		PSNR ↑	SSIM ↑	mIoU ↑	Acc ↑	rel ↓	τ ↑
2 input views							
Feature3DGS [73]	None	24.49	0.8132	0.4223	0.7174	12.95	21.07
LSM [15]	0.6B / 0.6B	24.39	0.8072	0.5078	0.7686	3.38	67.77
Uni3R [53]	1.3B / 1B	25.53	0.8727	0.5584	0.8268	3.87	61.37
DINOv3D	0.4B / 0.1B	**25.67**	**0.8823**	**0.5768**	**0.8512**	**3.02**	**70.52**
8 input views							
Feature3DGS [73]	None	18.17	0.674	0.195	0.724	17.28	13.31
Uni3R [53]	1.3B / 1B	24.12	0.829	0.554	0.807	4.46	56.88
DINOv3D	0.4B / 0.1B	**25.43**	**0.842**	**0.563**	**0.8325**	**3.38**	**66.74**
16 input views							
Feature3DGS [73]	None	17.09	0.649	0.198	0.672	23.71	10.57
Uni3R [53]	1.3B / 1B	23.24	0.8081	0.522	0.795	5.88	42.88
DINOv3D	0.4B / 0.1B	**24.68**	**0.823**	**0.541**	**0.8286**	**3.59**	**64.28**
32 input views							
Feature3DGS [73]	None	16.53	0.6127	0.176	0.621	28.95	7.89
Uni3R [53]	1.3B / 1B	22.54	0.7863	0.505	0.772	6.21	37.52
DINOv3D	0.4B / 0.1B	**24.32**	**0.802**	**0.529**	**0.8131**	**3.86**	**60.15**

heads. The network is fine-tuned for 100 epochs with a batch size of 1. To facilitate long-context learning across multiple views, sliding window attention is employed with a window size of 16 and a sequence length of 128.

Linear Probing Evaluation. Following the standard linear probing protocol established for DINOv3, the quality of the learned representations is evaluated by training linear layers on top of the frozen backbone.

4.3 Results on Unified 3D Scene Understanding

Table 1 reports the quantitative performance on ScanNet across varying input views. DINOv3D consistently establishes new state-of-the-art results across geometry, appearance, and semantics, demonstrating remarkable scalability and efficiency. Most notably, DINOv3D outperforms the billion-parameter Uni3R [53] while using only **10%** of the trainable parameters (0.1B *vs* 1.0B), validating the effectiveness of our progressive collaborative strategy.

Novel View Synthesis. As demonstrated in Table 1 and Fig. 3, DINOv3D achieves superior rendering fidelity when compared to both the large-scale LSM [15] and the unified framework of Uni3R [53]. Notably, under the sparse 2-view setting, DINOv3D attains a PSNR of **25.67 dB**, which surpasses the performance of Uni3R (25.53 dB) and LSM (24.39 dB). Furthermore, this performance advantage consistently persists as the number of input views increases. These results validate that the progressive collaborative strategy, which effectively adapts a frozen backbone rather than training from scratch, yields a highly robust and accurate 3D representation for novel view synthesis.

Fig. 3: Qualitative comparison of novel view synthesis on ScanNet (§4.4.3).

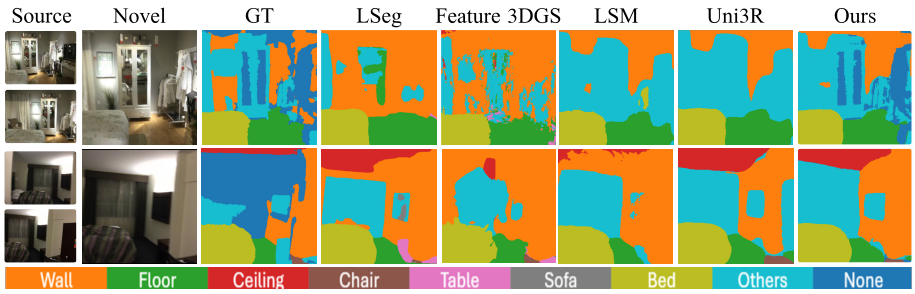

Fig. 4: Qualitative comparison of open-vocabulary semantic segmentation on ScanNet (§4.4.3). The dense 2D GT labels are obtained by projecting sparse 3D annotations, so some regions are inevitably incomplete or missing.

Open-Vocabulary Semantic Segmentation. As qualitatively illustrated in Fig. 4, the proposed ray-depth-semantic alignment yields clear benefits. Under the 2-view setting, DINOv3D achieves an **mIoU of 57.68%** and an **Accuracy of 85.12%**, significantly outperforming Uni3R and Feature3DGS. While baselines struggle with multi-view semantic inconsistency—evidenced by the performance decline of Feature3DGS with additional views—DINOv3D maintains robust consistency. This confirms that rectifying 2D features via geometric constraints effectively prevents semantic blurring in 3D space.

Depth Estimation. As illustrated in Fig. 5, DINOv3D establishes a new state-of-the-art for geometric reconstruction. Under the 2-view setting, DINOv3D achieves the lowest absolute relative error (**3.02**) and the highest $\delta < 1.25$ accuracy (**70.52%**). Compared to Uni3R (3.87) and LSM (3.38), DINOv3D recovers significantly more precise geometry. This enhanced precision is directly attributable to the integration of robust 3D priors, which effectively resolves ambiguities in texture-less regions where baseline methods fail.

4.4 Results on Geometry-specific Tasks

Geometry Reconstruction. Table 2 evaluates geometry reconstruction via novel view synthesis on RealEstate10K [74] under varying overlap settings. We compare DINOv3D against pose-required and pose-free baselines. Among the

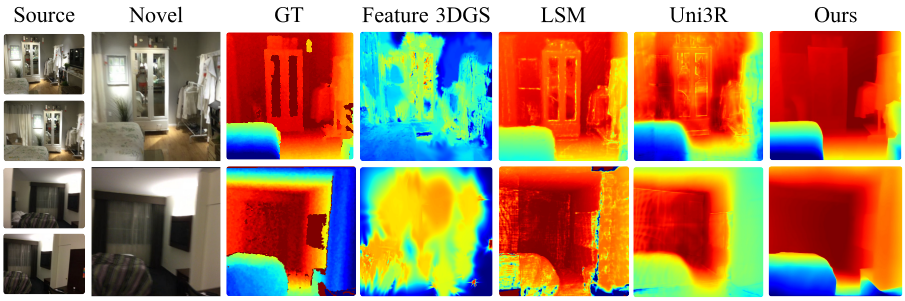

Fig. 5: Qualitative comparison of depth estimation on ScanNet (§4.3).

Table 2: Quantitative comparison (§4.4) of novel view synthesis on RealEstate10K [74] under various overlap settings.

Methods	Small			Medium			Large		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Pose-Required									
pixelSplat [7]	20.277	0.719	0.265	23.726	0.811	0.180	27.152	0.880	0.121
MVSplat [11]	20.371	0.725	0.250	23.808	0.814	0.172	27.466	0.885	0.115
Pose-Free									
MASt3R [31]	16.305	0.516	0.451	18.106	0.561	0.377	17.975	0.524	0.402
CoPoNeRF [22]	17.393	0.585	0.462	18.813	0.616	0.392	20.464	0.652	0.318
Splatt3R [51]	17.789	0.582	0.375	18.828	0.607	0.330	19.243	0.593	0.317
NoPoSplat [64]	22.514	0.784	0.210	24.899	0.839	0.160	27.411	0.883	0.119
SelfSplat [28]	14.828	0.543	0.469	18.857	0.679	0.328	23.338	0.798	0.208
DINOv3D	**23.585**	**0.812**	**0.186**	**25.869**	**0.861**	**0.142**	**28.231**	**0.903**	**0.102**

Table 3: Quantitative comparison (§4.4) on RealEstate10K [74] and ACID [36] for relative pose estimation.

Methods	RealEstate10K			ACID		
	5° ↑	10° ↑	20° ↑	5° ↑	10° ↑	20° ↑
DUSt3R [59]	0.329	0.537	0.691	0.113	0.273	0.469
MASt3R [31]	0.351	0.557	0.701	0.159	0.362	0.524
NoPoSplat [64]	0.568	0.737	0.839	0.342	0.504	0.653
SelfSplat [28]	0.223	0.413	0.589	0.213	0.372	0.541
DINOv3D	**0.635**	**0.774**	**0.861**	**0.371**	**0.544**	**0.687**

pose-free approaches, DINOv3D consistently achieves the best results across all metrics, significantly outperforming the strongest baseline, NoPoSplat [64]. Furthermore, it surpasses the performance of heavily optimized pose-required models, such as pixelSplat [7] and MVSplat [11]. For example, under the large overlap setting, DINOv3D attains a PSNR of **28.231**.

Pose Estimation. As detailed in Table 3, we evaluate relative pose estimation on RealEstate10K [74] and ACID [36]. We compare DINOv3D against recent baselines, including DUSt3R [59], MASt3R [31], NoPoSplat [64], and SelfSplat [28], using accuracy metrics under angular error thresholds of 5°, 10°, and 20°. The results demonstrate that our method consistently achieves the highest accuracy across all thresholds. Specifically, it yields a substantial improvement over the most competitive baseline, NoPoSplat [64]. For instance, under the strin-

Table 4: Quantitative comparison (§4.4) on ACID [36] and DTU [26] for cross-dataset generalization.

Methods	ACID			DTU		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
pixelSplat [7]	25.477	0.770	0.207	15.067	0.539	0.341
MVSplat [11]	25.525	0.773	0.199	14.542	0.537	0.324
NoPoSplat [64]	25.765	0.776	0.199	17.899	0.629	0.279
SelfSplat [28]	22.204	0.686	0.316	13.249	0.434	0.441
DINOv3D	**26.214**	**0.801**	**0.172**	**18.115**	**0.659**	**0.237**

Table 5: Linear probing evaluation on 2D tasks (§4.5). We compare our fine-tuned DINOv3D against the original DINOv3 baseline across diverse benchmarks.

Methods	Global Tasks				Dense Tasks				
	IN-ReaL↑	IN-R↑	Obj.Net↑	Ox.-H↑	ADE20k↑	NYUv2↓	DAVIS↑	NAVI↑	SPair↑
DINOv3	90.2	88.1	74.8	63.1	54.9	0.352	79.9	62.3	61.3
DINOv3D	**90.3**	**89.4**	**76.3**	**63.8**	**55.4**	**0.324**	**79.9**	**63.9**	**62.1**

gent 5° threshold, DINOv3D attains an accuracy of **0.635** on RealEstate10K and **0.371** on ACID. These findings validate the superiority of DINOv3D in accurately recovering relative camera poses.

Cross-dataset Generalization. As presented in Table 4, we evaluate the cross-dataset generalization on ACID [36] and DTU [26]. We compare DINOv3D against recent baselines, including pixelSplat [7], MVSplat [11], NoPoSplat [64], and SelfSplat [28], using standard image synthesis metrics. The quantitative results demonstrate that our method consistently achieves the highest performance across all metrics. Notably, on ACID, DINOv3D attains a PSNR of **26.214** and an SSIM of **0.801**, outperforming the most competitive baseline, NoPoSplat [64]. Furthermore, on DTU, DINOv3D maintains superior robustness despite the domain gap, yielding a PSNR of **18.115** and an LPIPS of **0.237**.

4.5 Results on 2D Linear Probing

To verify 3D reconstruction as an effective regularizer for 2D representation learning, we evaluate the fine-tuned backbone features via linear probing, following the same protocol in [50]. As shown in Table 5, DINOv3D consistently outperforms the DINOv3 baseline across global classification and dense prediction tasks, mitigating the catastrophic forgetting typically associated with domain-specific fine-tuning. Substantial improvements occur in tasks requiring spatial and geometric reasoning. For instance, depth estimation RMSE on NYUv2 [49] decreases to **0.324**, indicating enhanced metric scale awareness. Furthermore, accuracy on ObjectNet [1] improves by **1.5%**, while point correspondence performance on NAVI [25] and SPair [40] increases by **1.6%** and **0.8%**, respectively. These gains demonstrate that multi-view consistency constraints facilitate learning view-invariant spatial representations. Beyond geometric tasks, DINOv3D exhibits enhanced robustness on standard classification benchmarks.

Table 6: Ablation of key components on ScanNet and ADE20k (§4.6). We incrementally add the Dual Branch architecture (Dual Br.), Long Context (128 frames), Ray-Depth-Semantic Alignment (Align.), and the Uni3D Dataset to the baseline.

Methods	Dual Br.	Align.	Long Ctx.	ScanNet (3D)				ADE20k (2D)
				PSNR↑	SSIM↑	rel↓	mIoU↑	mIoU↑
Baseline (Frozen)	-	-	-	23.45	0.782	3.85	51.2	54.9
+ Dual Br.	✓	-	-	24.62	0.825	3.42	54.5	55.0
+ Align.	✓	✓	-	25.41	0.868	3.15	55.6	55.2
+ Long Context	✓	✓	✓	25.58	0.877	3.11	56.5	55.3
+ Uni3D (Ours)	✓	✓	✓	**25.67**	**0.882**	**3.02**	**57.7**	**55.4**

Table 7: Ablation of input context length on ScanNet (§4.6).

Context Length	8 frames	32 frames	64 frames	**128 frames**	144 frames
PSNR ↑	24.62	25.10	25.45	**25.67**	25.69
rel ↓	3.42	3.28	3.11	**3.02**	3.01

Table 8: Ablation of fine-tuning strategies on ScanNet and NYUv2 (§4.6).

Strategy	3D Task (ScanNet)		2D Task (NYUv2)	
	PSNR ↑	rel ↓	RMSE ↓	$mAcc$ ↑
Frozen Backbone	24.15	3.65	0.352	81.2
Full Fine-tuning	**25.72**	**3.01**	0.385	77.5
LoRA (Ours)	25.67	3.02	**0.324**	**84.6**

4.6 Ablation Studies

Effectiveness of Key Components. Table 6 presents an ablation study of the proposed components. The baseline yields a PSNR of 23.45 dB on ScanNet dataset and an mIoU of 54.9% on ADE20K dataset. Integrating the dual-branch architecture reduces the relative error to 3.42 and effectively improves two-dimensional knowledge retention. Subsequently, adding ray-depth-semantic alignment and extending the context length to 128 frames progressively enhance three-dimensional geometric metrics. Finally, training on the proposed Uni3D dataset achieves optimal performance, notably boosting the mIoU on ScanNet dataset to **57.7%**, the PSNR to **25.67 dB**, and the mIoU on ADE20K dataset to **55.4%**. These results validate the efficacy of the whole design.

Impact of Context Length. Table 7 evaluates the effect of varying the input sequence length during training. Extending the context from 8 to 128 frames consistently enhances the reconstruction quality. The performance saturates at 128 frames, indicating an optimal balance between capturing spatial-temporal consistency and maintaining computational efficiency.

Fine-tuning Strategy: Frozen *vs* Full *vs* LoRA. Table 8 evaluates strategies for updating the backbone. While a **Frozen** backbone restricts 3D adaptation, **Full Fine-tuning** achieves strong 3D performance at the cost of catastrophic forgetting on 2D tasks. The proposed **LoRA-based** approach strikes an optimal balance, maintaining competitive 3D reconstruction quality while simultaneously enhancing the 2D representations (yielding the lowest NYUv2 RMSE of 0.324). This validates the efficacy of the collaborative optimization paradigm.

5 Conclusion

In this paper, we presented DINOv3D, a joint optimization framework that unites the paradigms of 2D and 3D spatial understanding through 3D Gaussian Splatting. By employing a homologous teacher-student architecture, DINOv3D effectively absorbs multi-view geometric constraints while preventing the catastrophic forgetting of pre-trained 2D features. Extensive experiments demonstrate that DINOv3D establishes state-of-the-art performance across diverse and challenging 2D&3D spatial perception tasks.

Acknowledgements

This work was supported by Zhejiang Provincial Natural Science Foundation of China (No. LR26F020002), National Natural Science Foundation of China (No. 62372405), State Key Laboratory of Autonomous Intelligent Unmanned Systems and National Defense Science and Technology Industry Bureau Technology Infrastructure Project (JSZL2024606C001).

References

1. Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B.: Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. NeurIPS (2019)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: ICCV (2021)
3. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., Oord, A.v.d.: Are we done with imagenet? arXiv preprint arXiv:2006.07159 (2020)
4. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: Must3r: Multi-view Network for Stereo 3d Reconstruction. In: CVPR (2025)
5. Cai, X., Pei, G., Sun, Z., Yao, Y., Shen, F., Wang, W.: Iris: Bringing real-world priors into diffusion model for monocular depth estimation. In: CVPR (2026)
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Surís, D., Ryali, C.K., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Radle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H., Doll'ar, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment Anything with Concepts. In: ICLR (2026)
7. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: CVPR (2024)
8. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: ICCV (2021)
9. Chen, G., Wang, W.: A survey on 3d gaussian splatting. ACM Computing Surveys (2026)
10. Chen, M., Chen, G., Wang, W., Yang, Y.: Hydra-sgg: Hybrid relation assignment for one-stage scene graph generation. In: ICLR (2025)

11. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In: ECCV (2024)
12. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
13. Ding, X., Gao, J., Pan, C., Wang, W., Qin, J.: History-enhanced two-stage transformer for aerial vision-and-language navigation. In: AAAI (2026)
14. Fan, S., Liu, R., Wang, W., Yang, Y.: Scene map-based prompt tuning for navigation instruction generation. In: CVPR (2025)
15. Fan, Z., Zhang, J., Cong, W., Wang, P., Li, R., Wen, K., Zhou, S., Kadambi, A., Wang, Z., Xu, D., Ivanovic, B., Pavone, M., Wang, Y.: Large Spatial Model: End-to-end Unposed Images to Semantic 3d. In: NeurIPS (2024)
16. Feng, T., Wang, W., Yang, Y.: Gaussian-based world model: Gaussian priors for voxel-based occupancy prediction and future motion prediction. In: CVPR (2025)
17. Gao, J., Liu, R., Wang, W.: 3d gaussian map with open-set semantic grouping for vision-language navigation. In: ICCV (2025)
18. Gao, J., Liu, R., Xu, Y., Cao, T., Zhang, Y., Zhang, Z., Peng, S., Yang, Y., Wang, W.: Uncertainty-aware gaussian map for vision-language navigation. In: ICLR (2026)
19. Gao, J., Wang, C., Zhang, W., Li, J., Li, L.A., Wang, W., Zhu, Y., Wang, Y.: Clinically-grounded counterfactual reasoning for medical video diagnosis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7014–7025 (2026)
20. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: COLM (2024)
21. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: ICCV (2021)
22. Hong, S., Jung, J., Shin, H., Yang, J., Kim, S., Luo, C.: Unifying correspondence pose and nerf for generalized pose-free novel view synthesis. In: CVPR (2024)
23. Hou, J., Xie, S., Graham, B., Dai, A., Nießner, M.: Pri3d: Can 3d priors help 2d representation learning? In: CVPR (2021)
24. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR (2022)
25. Jampani, V., Maninis, K.K., Engelhardt, A., Karpur, A., Truong, K., Sargent, K., Popov, S., Araujo, A., Martin Brualla, R., Patel, K., et al.: Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. NeurIPS (2023)
26. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: CVPR (2014)
27. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., Lin, D., Dai, B.: Anysplat: Feed-forward 3d Gaussian Splatting from Unconstrained Views. TOG (2025)
28. Kang, G., Yoo, J., Park, J., Nam, S., Im, H., Shin, S., Kim, S., Park, E.: Selfsplat: Pose-Free and 3d Prior-Free Generalizable 3d Gaussian Splatting. In: CVPR (2025)
29. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM TOG (2023)
30. Krakovsky, S., Fiebelman, G., Benaim, S., Averbuch-Elor, H.: Lang3d-XL: Language Embedded 3d Gaussians for Large-scale Scenes. In: Proceedings of the SIGGRAPH Asia Conference Papers. pp. 1–11. ACM (2025)
31. Leroy, V., Cabon, Y., Revaud, J.: Grounding Image Matching in 3d with MASt3R. In: ECCV (2025)

32. Li, B., Weinberger, K.Q., Belongie, S.J., Koltun, V., Ranftl, R.: Language-driven Semantic Segmentation. In: ICLR (2022)
33. Li, Q., Sun, J., An, L., Su, Z., Zhang, H., Liu, Y.: Semanticsplat: Feed-Forward 3d Scene Understanding with Language-Aware Gaussian Fields. arXiv preprint arXiv:2506.09565 (2025)
34. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. In: ICLR (2026)
35. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: CVPR (2024)
36. Liu, A., Tucker, R., Jampani, V., Makadia, A., Snavely, N., Kanazawa, A.: Infinite nature: Perpetual view generation of natural scenes from a single image. In: ICCV (2021)
37. Liu, R., Duan, Z., Gao, J., Yang, Y., Wang, W.: Uncertainty-aware 3d reconstruction for dynamic underwater scenes (2026)
38. Liu, Y., Fan, K., Yu, W., Li, C., Lu, H., Yuan, Y.: Monosplat: Generalizable 3d Gaussian Splatting from Monocular Depth Foundation Models. In: CVPR (2025)
39. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
40. Min, J., Lee, J., Ponce, J., Cho, M.: Spair-71k: A large-scale benchmark for semantic correspondence. arXiv preprint arXiv:1908.10543 (2019)
41. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. TMLR (2024)
42. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: CVPR (2023)
43. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. In: CVPRW (2017)
44. Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: CVPR (2018)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICLR (2021)
46. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: ICLR (2025)
47. Segre, L., Hirschorn, O., Avidan, S.: Multi-View Foundation Models. arXiv.org **abs/2512.15708** (2025)
48. Sheng, Y., Deng, J., Zhang, X., Zhang, Y., Hua, B., Zhang, Y., Ji, J.: Spatialsplat: Efficient Semantic 3d from Sparse Unposed Images. In: ICCV (2025)
49. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
50. Sim'eoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., J'egou, H., Labatut, P., Bojanowski, P.: Dinov3. arXiv (2025)
51. Smart, B., Zheng, C., Laina, I., Prisacariu, V.: Splatt3r: Zero-shot Gaussian Splatting from Uncalibrated Image Pairs. arXiv preprint arXiv:2408.13912 (2024)

52. Sun, W., Zhou, Y., Jiao, J., Li, Y.: Cags: Open-vocabulary 3d scene understanding with context-aware gaussian splatting. arXiv preprint arXiv:2504.11893 (2025)
53. Sun, X., Jiang, H., Liu, L., Nam, S., Kang, G., Wang, X., Sui, W., Su, Z., Liu, W., Wang, X., Park, E.: Uni3r: Unified 3d reconstruction and semantic understanding via generalizable gaussian splatting from unposed multi-view images. In: CVPR (2026)
54. Sun, Z., Yao, Y., Liu, T., Li, Z., Shen, F., Tang, J.: Jo-snc: Combating noisy labels through fostering self-and neighbor-consistency. TPAMI (2025)
55. Tian, Q., Tan, X., Gong, J., Xie, Y., Ma, L.: Uniforward: Unified 3d scene and semantic field reconstruction via feed-forward gaussian splatting from only sparse-view images. arXiv preprint arXiv:2506.09378 (2025)
56. Tian, Q., Tan, X., Ying, J., Wang, X., Xie, Y., Ma, L.: Fleg: Feed-forward language embedded gaussian splatting from any views. arXiv preprint arXiv:2512.17541 (2025)
57. Venkataramanan, S., Rizve, M.N., Carreira, J., Asano, Y.M., Avrithis, Y.: Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video. In: ICLR (2024)
58. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotný, D.: Vggt: Visual Geometry Grounded Transformer. In: CVPR (2025)
59. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d Vision Made Easy. In: CVPR (2024)
60. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depthsplat: Connecting Gaussian Splatting and Depth. In: CVPR (2024)
61. Yang, Z., Pei, G., Chen, T., Yuan, X., Zhang, H., Shu, X., Yao, Y.: Beyond quadratic: Linear-time change detection with rwkv. In: AAAI (2026)
62. Yang, Z., Pei, G., Chen, T., Zhou, Y., Zhou, T., Yao, Y., Shen, F.: Efficiency follows global-local decoupling. In: CVPR (2026)
63. Yang, Z., Pei, G., Yao, Y., Zhou, T., Ding, L., Shen, F.: ChangeTitans: Toward remote sensing change detection with neural memory. TGRS (2025)
64. Ye, B., Liu, S., Peng, S., Xu, H., Li, X., Yang, M.H., Pollefeys, M.: No Pose, No Problem: Surprisingly Simple 3d Gaussian Splats from Sparse Unposed Images. In: ICLR (2025)
65. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: ICCV (2023)
66. Yin, J., Jiang, X., Chen, T., Pei, G., Yao, Y., Shen, F., Shen, H.T.: Depmatch: Boosting semi-supervised semantic segmentation by exploring depth difference knowledge. TIP (2026)
67. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: CVPR (2021)
68. Yue, Y., Das, A., Engelmann, F., Tang, S., Lenssen, J.E.: Improving 2d feature representations by 3d-aware fine-tuning. In: ECCV (2024)
69. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
70. Zhou, B., Lai, Q., Sun, Z., Shu, X., Yao, Y., Wang, W.: Learning 3d representations for spatial intelligence from unposed multi-view images. In: CVPR (2026)
71. Zhou, B., Li, L., Wang, Y., Liu, H., Yao, Y., Wang, W.: Unialign: Scaling multimodal alignment within one unified model. In: CVPR (2025)
72. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)

73. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d Gaussian Splatting to Enable Distilled Feature Fields. In: CVPR (2024)
74. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. TOG **37**(4), 65 (2018)
75. Zhu, R., Hui, K.H., Liu, Z., Wu, Q., Tang, W., Qiu, S., Heng, P.A., Fu, C.W.: Cos3d: Collaborative open-vocabulary 3d segmentation. In: NeurIPS (2025)