This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

E-RayZer: Self-supervised 3D Reconstruction as Spatial Visual Pre-training

Qitao Zhao[1] Hao Tan[2] Qianqian Wang[3] Sai Bi[2]
Kai Zhang[2] Kalyan Sunkavalli[2] Shubham Tulsiani[1*] Hanwen Jiang[2*]
[1]Carnegie Mellon University [2]Adobe Research [3]Harvard University [*]Equal advising
Project & Code: qitaozhao.github.io/E-RayZer

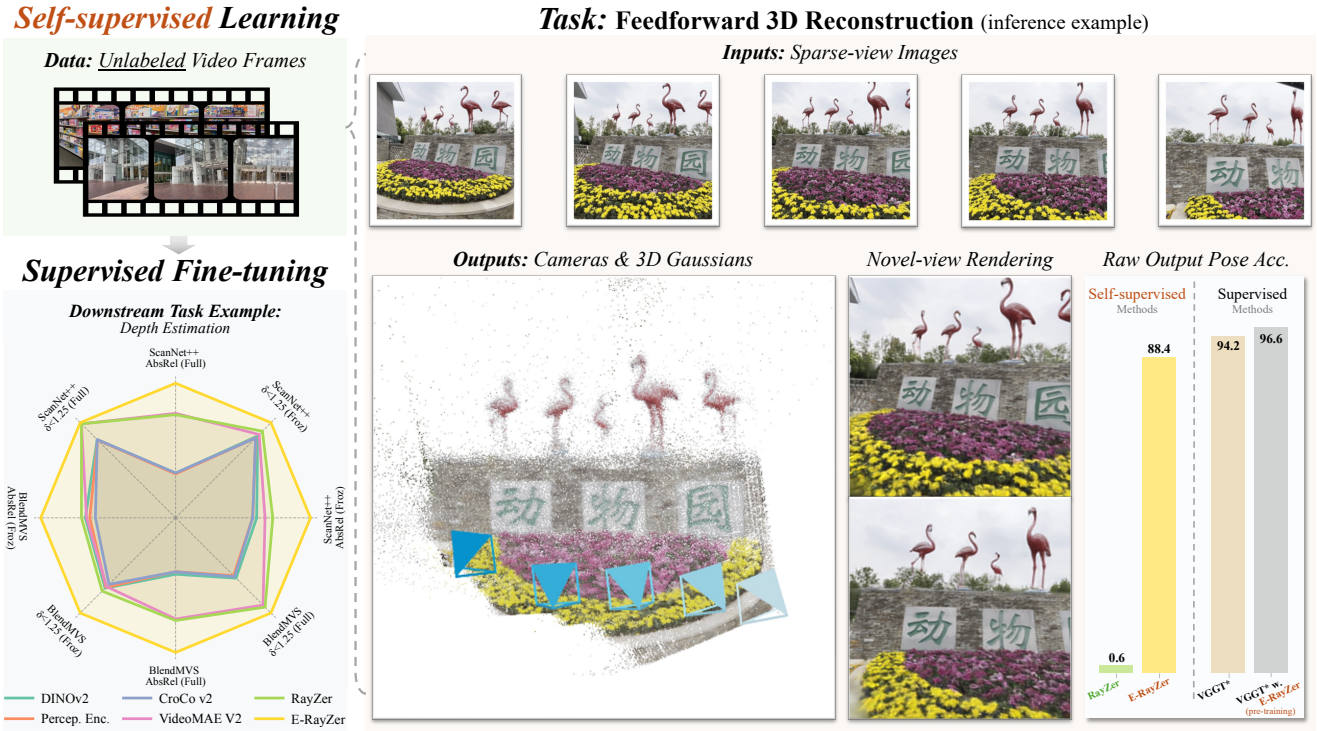

Figure 1. **E-RayZer**, a **self-supervised** 3D vision model that predicts camera poses and scene geometry as 3D Gaussians. The use of **explicit** 3D geometry yields more geometrically grounded poses compared to its implicit counterpart, RayZer [25]: they are comparable to (and sometimes surpass) those from our supervised baseline, VGGT [58]. Furthermore, E-RayZer serves as a self-supervised **visual pre-training** framework, with learned representations that transfer effectively to downstream tasks requiring 3D understanding, outperforming previous representation learners such as CroCo v2 [64], VideoMAE V2 [59], DINOv3 [50], and Perception Encoder [7].

Abstract

*Self-supervised pre-training has driven rapid progress in foundation models for language, 2D images, and video, yet remains largely unexplored for learning 3D-aware representations from multi-view images. In this paper, we present E-RayZer, a self-supervised 3D vision model that learns geometrically grounded representations directly from unlabeled images. Unlike prior self-supervised methods such as RayZer, which infer 3D indirectly through latent-space view synthesis, **E**-RayZer operates directly in 3D space, performing self-supervised 3D reconstruction with **Explicit** geometry. This formulation eliminates shortcut solutions and yields representations that are 3D-aware. To ensure convergence and scalability, we introduce a fine-grained learning curriculum that organizes training from easy to hard samples and harmonizes heterogeneous data sources without any supervision. Experiments show that E-RayZer significantly outperforms RayZer on pose estimation and matches or sometimes surpasses fully supervised reconstruction models such as VGGT[1]. Furthermore, its learned representations outperform leading visual pre-training models (*e.g.*, DINOv3, CroCo v2, VideoMAE V2, and RayZer) on 3D downstream tasks, establishing E-RayZer as a promising paradigm for spatial visual pre-training.*

[1]We reproduce VGGT with matched model architecture and training setups, and denote it by VGGT* as a supervised baseline.

1. Introduction

Pre-training with self-supervision forms the foundation of frontier models, enabling them to learn *meaningful representations* from vast amounts of unlabeled data. This paradigm has proven effective for text [8, 13], 2D images [19, 38], and video [2, 53], where large models capture language semantics, visual concepts, and temporal dynamics. However, we argue that one essential component is still missing – **learning 3D-aware representations from unlabeled multi-view images** – as 3D spatial understanding is fundamental for perceiving and interacting with the physical world. Yet current 3D vision models mostly follow a different route: *fully supervised learning* using 3D pseudo-labels estimated by SfM systems (*e.g.*, COLMAP [47]), which is inherently inefficient, imperfect, and ultimately unscalable. To move forward, we need a self-supervised pre-training framework that can learn 3D-aware representations from abundant raw visual observations.

In this paper, we present **E-RayZer**, the first **truly self-supervised 3D Gaussian splatting reconstruction** model that learns 3D-aware representations from unlabeled data, establishing a new paradigm for *3D spatial visual pre-training* (Fig. 1). Unlike its predecessor RayZer [25], which exhibits only *superficial* 3D awareness through the *proxy task* of self-supervised view synthesis in *latent space*, E-RayZer operates directly in *3D space*, learning self-supervised 3D reconstruction. Concretely, E-RayZer predicts camera parameters and 3D Gaussians [30] from inputs and renders them back for photometric self-supervision under physical rendering constraints. By grounding representations in **explicit** scene geometry, E-RayZer learns features that are genuinely 3D-aware and free from RayZer's shortcut solutions such as frame interpolation (see Sec. 3.1). This design yields both a *camera space* that is more geometrically grounded and interpretable than RayZer's, and *latent representations* that are truly **3D-aware**, effectively benefiting downstream 3D vision tasks.

Although explicit 3D Gaussians offer clear advantages, they also introduce substantial training challenges. As reported in RayZer (Tab. 7), training with explicit 3D leads to non-convergence. To address this, we propose a **fine-grained learning curriculum** built on the concept of *visual overlap* between input views. We begin with samples of *high visual overlap*, allowing the pose estimator to initialize from near-identity poses, and gradually reduce overlap to promote general 3D understanding. When *scaling* to heterogeneous training sources, visual overlap provides a natural and unified metric to adaptively align varying camera motion distributions, improving data consistency. Notably, we approximate visual overlap in an *unsupervised* manner, keeping the framework entirely free from 3D annotations.

We systematically study E-RayZer's performance across different training data scales. We highlight key conclusions and summarize our contributions as follows:

- E-RayZer is the first **truly self-supervised feed-forward 3D Gaussian splatting reconstruction** model, trained **from scratch** with zero 3D annotation.
- E-RayZer **outperforms prior visual representation learners**, *e.g.*, DINOv3 [50], CroCo v2 [64], VideoMAE V2 [59], and Perception Encoder [7], on downstream 3D tasks (Tab. 3-4), establishing E-RayZer as a strong paradigm for *spatial visual pre-training*.
- Compared with previous self-supervised 3D vision models, E-RayZer exhibits **stronger 3D understanding**, as evidenced by significantly improved unsupervised camera pose estimation (Tab. 1) and the fine-tuning results on downstream 3D tasks (Tab. 3).
- Compared with state-of-the-art supervised models, *e.g.*, VGGT [58] (reproduced with matched architecture and training setups), E-RayZer achieves **comparable or sometimes superior performance** (Tab. 2) and exhibits *similar scaling behavior* (Tab. 5), despite being purely self-supervised.

2. Related Work

Supervised Pose Estimation and 3D Reconstruction. Early learning-based methods estimated relative camera poses from image pairs [3, 4, 9, 44], while later approaches extended this to multi-view reasoning across multiple inputs [23, 24, 33, 51, 57, 76, 77]. Given posed images, 3D representations can be reconstructed either by direct regression [23, 74, 78] or by optimization-based mode seeking with diffusion models [80, 84]. More recently, several methods have unified pose estimation and 3D reconstruction by predicting pixel-aligned pointmaps [14, 58, 60, 62, 81]; these exhibit strong robustness under sparse inputs and generalize well across diverse domains [55]. Nevertheless, training such supervised models still relies on camera pose and dense depth annotations, typically obtained from traditional SfM systems (*e.g.*, COLMAP [47]), which can be inaccurate and ultimately limit performance.

Recent work has also investigated predicting 3D Gaussians [30] using photometric losses as (part of the) supervision. However, these methods remain *de facto* supervised by 3D annotations, as they rely on ground-truth intrinsics [20, 28, 70] and/or target-view camera poses during training [28, 52, 70], or require initialization and/or regularization from 3D-supervised models [21, 26, 52]. In contrast, E-RayZer requires no 3D supervision and can be trained from scratch, making it *truly self-supervised* – while achieving even stronger performance.

Self-supervised Novel-view Synthesis. To alleviate the dependence on 3D supervision, another line of research learns scene representations directly from 2D images via novel-view synthesis. Early work predicted scene features from a single viewpoint and rendered target views as su-

pervision [17, 31, 65, 82]. More recently, RUST [45], RayZer [25], and others [37, 56, 61] adopt learning-based latent rendering from multi-view inputs. However, these methods exhibit limited 3D awareness. For instance, RayZer learns view interpolation within an uninterpretable pose space. We build on RayZer but differ in three key respects: we adopt an explicit 3D representation (*i.e.*, 3D Gaussians [30]), a more principled learning curriculum, and larger-scale training. We show that explicit 3D modeling yields more geometrically grounded representations, establishing a promising pre-training framework for downstream tasks that require 3D understanding.

Closely relevant to our work, DBARF [12] explored a NeRF-based [36] framework for self-supervised pose estimation and novel-view synthesis using a multi-stage pipeline. E-RayZer offers a more streamlined and scalable framework, and goes beyond novel-view synthesis by investigating the learned representations for downstream tasks.

Visual Pre-training for Representation Learning. Prior work has made substantial progress in learning global image semantics through image-language association [1, 41, 54], self-distillation [10, 38, 50], contrastive learning [11, 18], masked image modeling [19, 39], and video-level temporal self-supervision [5, 15, 53]. However, learning *3D-aware and geometrically grounded representations* remains underexplored, despite its potential to benefit 3D-related tasks where supervision is scarce. Recent efforts explore learning 3D awareness through proxy tasks such as latent-space novel-view synthesis [25, 63, 64], but whether they enforce true 3D understanding remains unclear. In this work, E-RayZer addresses this gap with explicit 3D modeling and a learning curriculum that enables effective scaling, yielding 3D-grounded representations.

3. Approach

From unlabeled multi-view images, E-RayZer learns to predict camera parameters (poses and intrinsics) and **explicit** 3D scene geometry under self-supervision. Its learned representations can be further leveraged for downstream tasks, demonstrating E-RayZer's potential as a 3D-aware visual pre-training framework.

In the following, we first revisit RayZer [25], the *implicit* predecessor, and discuss its limitations (Sec. 3.1). Building on RayZer's core design while addressing these issues by leveraging *Explicit* 3D modeling, we introduce *E*-RayZer (Sec. 3.2). Finally, we present a sequence-level curriculum learning strategy based on visual overlap between frames to improve performance and scalability (Sec. 3.3).

3.1. Preliminaries: RayZer with Implicit 3D

RayZer splits all input images into two *non-overlapping* subsets: an "observed" reference set ($\mathcal{I}_{\text{ref}}$) for latent scene inference, and a "hidden" target set ($\mathcal{I}_{\text{tgt}}$) for providing self-supervision. RayZer uses predicted cameras of target views ($\mathcal{I}_{\text{tgt}}$) to render the scene predicted from the reference views ($\mathcal{I}_{\text{ref}}$), and applies the photometric loss as self-supervision:

$$\mathcal{L} = \Sigma_{(I,\hat{I})\in(\mathcal{I}_{\text{tgt}},\hat{\mathcal{I}}_{\text{tgt}})}\big(\texttt{MSE}(I,\hat{I}) + \lambda\cdot\texttt{Percep}(I,\hat{I})\big), \quad (1)$$

where `Percep` denotes the perceptual loss [27].

RayZer leverages transformers for pose estimation, latent (implicit) scene reconstruction, and rendering. It first predicts camera intrinsics and extrinsics for all input images $\mathcal{I}\in\mathbb{R}^{V\times H\times W\times 3}$ using a multi-view transformer $f_{\boldsymbol{\theta}}^{\text{cam}}$, as:

$$(\mathbf{K},\mathbf{T}) = f_{\boldsymbol{\theta}}^{\text{cam}}(\mathcal{I}), \quad \mathbf{T}_i = [\mathbf{R}_i \,|\, \mathbf{t}_i] \in SE(3), \quad (2)$$

where $\mathbf{K}\in\mathbb{R}^{3\times 3}$ is the intrinsics shared by all views, $\mathbf{T}\in\mathbb{R}^{V\times 4\times 4}$ denotes the extrinsics, and $i=1,\ldots,V$ indexes the input images. Each camera $(\mathbf{K},\mathbf{T}_i)$ is then converted into a pixel-aligned Plücker ray map $\mathbf{R}_i^{\text{plk}}$ [40, 77].

To infer latent scene representations, RayZer tokenizes the concatenation (along the feature dimension) of image and rays for $\mathcal{I}_{\text{ref}}$ and updates a set of learnable scene tokens $\mathbf{z}_0^{\text{scene}}$ through another transformer $f_{\boldsymbol{\psi}}^{\text{scene}}$, as:

$$\mathbf{z}_{\text{ref}}^{\text{scene}} = f_{\boldsymbol{\psi}}^{\text{scene}}\big(\mathbf{z}_0^{\text{scene}}, \text{Linear}(\mathcal{I}_{\text{ref}}, \mathbf{R}_{\text{ref}}^{\text{plk}})\big), \quad (3)$$

where $\text{Linear}(\cdot)$ denotes a patch-wise linear projection for fusing and tokenizing RGB and ray information. The resulting $\mathbf{z}_{\text{ref}}^{\text{scene}}$ represents the latent scene features.

For rendering, the self-predicted target-view Plücker ray maps are likewise tokenized and concatenated with the scene representation $\mathbf{z}_{\text{ref}}^{\text{scene}}$ (along the token dimension). These target-view ray tokens are refined via transformer $f_{\boldsymbol{\phi}}^{\text{rend}}$ and finally decoded to RGB images, as:

$$\hat{\mathcal{I}}_{\text{tgt}} = f_{\boldsymbol{\phi}}^{\text{rend}}\big(\mathbf{z}_{\text{ref}}^{\text{scene}}, \text{Linear}(\mathbf{R}_{\text{tgt}}^{\text{plk}})\big). \quad (4)$$

Then RayZer applies photometric self-supervision (Eq. 1).

Limitations of RayZer's Implicit 3D. RayZer achieves high-fidelity novel-view synthesis. However, **RayZer is not fully 3D-grounded**. Since its camera estimation ($f_{\boldsymbol{\theta}}^{\text{cam}}$), latent scene reconstruction ($f_{\boldsymbol{\psi}}^{\text{scene}}$), and rendering ($f_{\boldsymbol{\phi}}^{\text{rend}}$) modules are jointly learned from scratch, they only need to remain *mutually compatible*, but are not guaranteed to be physically or spatially meaningful. This issue is further amplified by RayZer's pure transformer-based architecture, which contains almost *no 3D inductive bias* and thus possesses excessive flexibility to learn undesirable shortcut solutions. As evidenced by its imperfect camera pose distribution, RayZer relies on a mixture of true 3D understanding and video-interpolation priors to achieve high-quality synthesis. While this design suffices for novel-view synthesis, it limits RayZer's potential as a *spatial pre-training* framework for learning genuinely 3D-aware representations.

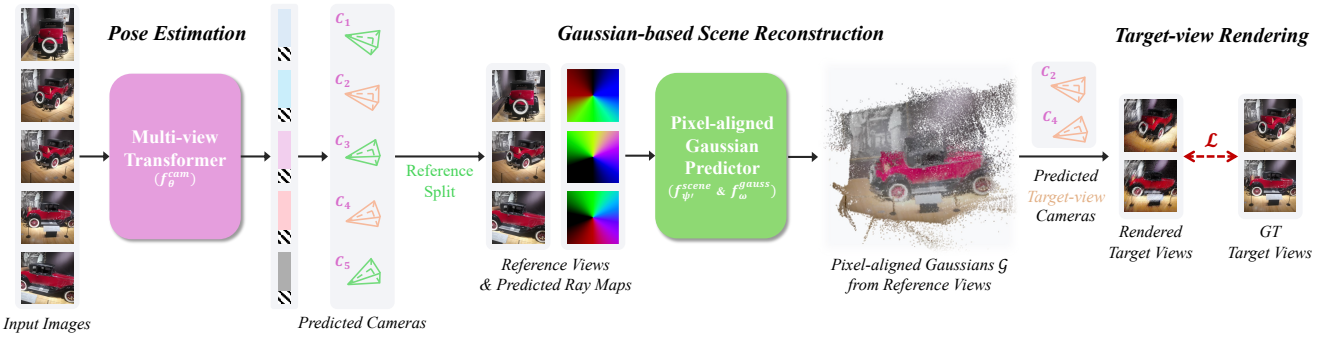

Figure 2. **E-RayZer Model & Training.** E-RayZer first predicts camera poses and intrinsics for all images. Then it follows RayZer [25] to split images into two sets. E-RayZer predicts explicit 3D Gaussians as scene representation from the reference views ($\mathcal{I}_{\text{ref}}$), and renders the scene using self-predicted target-view ($\mathcal{I}_{\text{tgt}}$) cameras. Finally, E-RayZer is trained with self-supervised photometric losses on target views.

3.2. E-RayZer: Explicit 3D with Self-supervision

Our Insights. We argue that **3D inductive biases remain essential** for 3D representation learning, but must be introduced in ways that preserve **learning scalability**.

We therefore inject a *lightweight* 3D inductive bias through model design while keeping training fully self-supervised, striking a balance between *3D awareness* and *scalability*. Specifically, E-RayZer replaces RayZer's latent scene representation with *explicit* 3D geometry (*i.e.*, 3D Gaussians [30]), providing geometric regularization that yields more grounded pose estimation, scene reconstruction, and latent representations.

Overview. As shown in Fig. 2, E-RayZer first predicts the camera parameters for all images, and then infers pixel-aligned 3D Gaussians $\mathcal{G}$ from the reference view subset ($\mathcal{I}_{\text{ref}}$). Then E-RayZer predicts the target view subset ($\mathcal{I}_{\text{tgt}}$), by rendering the 3D Gaussians predicted from $\mathcal{I}_{\text{ref}}$ under self-predicted cameras of $\mathcal{I}_{\text{tgt}}$. Since 3D Gaussians support closed-form differentiable rendering, the latent rendering decoder used in RayZer (*i.e.*, f^{rend}_{ϕ} in Eq. 4) is no longer required. We now describe our *key differences* from RayZer while elaborating on details.

Gaussian-based Scene Reconstruction. E-RayZer first predicts the cameras for all views in a similar way with RayZer (besides differences in model architecture that will be detailed later). Then, E-RayZer directly transforms the "posed" reference views to pixel-aligned 3D Gaussians. We first encode posed reference views into latent tokens:

$$\mathbf{s}_{\text{ref}} = f^{\text{scene}}_{\psi'}\big(\text{Linear}(\mathcal{I}_{\text{ref}}, \mathbf{R}^{\text{plk}}_{\text{ref}})\big) \tag{5}$$

where $\mathbf{s}_{\text{ref}} \in \mathbb{R}^{K_{\text{ref}}hw \times C}$ denotes the updated image tokens of reference views after multi-view aggregation. In detail, K_{ref} is the number of views in $\mathcal{I}_{\text{ref}}$, $h = H/p$ and $w = W/p$ are token number along height and width dimensions using a patch size of p, and C is channel dimension of the latent space. Note that the complexity of global attention in Eq. 5 is $\mathcal{O}((K_{\text{ref}}hw)^2)$, while it is $\mathcal{O}((K_{\text{ref}}hw+n_{\mathbf{z}})^2)$ for RayZer (Eq. 3), where $n_{\mathbf{z}}$ is the size for RayZer's scene token set.

Then, we use a lightweight decoder to transform the updated image tokens $\mathbf{s}_{\text{ref}}$ into per-pixel 3D Gaussian parameters along each camera ray across all reference views, as:

$$\begin{aligned} &\mathcal{G} = f^{\text{gauss}}_{\omega}(\mathbf{s}_{\text{ref}}), \quad \text{where} \\ &\mathcal{G} = \big\{\, g_i = (d_i,\, \mathbf{q}_i,\, \mathbf{C}_i,\, \mathbf{s}_i,\, \alpha_i) \,\big\}_{i=1}^{K_{\text{ref}}\times H\times W}. \end{aligned} \tag{6}$$

These parameters include the distance along the ray $d_i \in \mathbb{R}$, orientation represented as a quaternion $\mathbf{q}_i \in \mathbb{R}^4$, spherical harmonic coefficients $\mathbf{C}_i \in \mathbb{R}^{(d_{\text{SH}}+1)^2\times 3}$, scale $\mathbf{s}_i \in \mathbb{R}^3$, and opacity $\alpha_i \in \mathbb{R}$. The predicted 3D Gaussians collectively represent the scene geometry.

We then use E-RayZer's self-predicted target views cameras, denoted as $\mathcal{C}_{\text{tgt}} = \{\, (\mathbf{K}, \mathbf{T}_i) \mid i \in \mathcal{I}_{\text{tgt}} \,\}$, to render the 3D Gaussians $\mathcal{G}$ and get prediction of target views, as:

$$\hat{\mathcal{I}}_{\text{tgt}} = \pi(\mathcal{G}, \mathcal{C}_{\text{tgt}}), \tag{7}$$

where π denotes the differentiable rendering operations of 3D Gaussians. Note that we modify gsplat [72] to support gradient back-propagation to camera intrinsics $\mathbf{K}$. Compared with RayZer, this design improves both *rendering efficiency* and *3D-awareness* by removing the need to learn a transformer-based renderer. Finally, we apply photometric loss on the rendered target views as Eq. 1.

Avoiding Undesirable View Interpolation. As discussed in Sec. 3.1, RayZer tends to learn undesirable frame interpolation cues as shortcut solutions. We identify a main cause as its use of *image index embeddings* to associate image tokens with corresponding camera tokens for camera estimation, which provides a strong cue for learning interpolation.

In E-RayZer, we remove the image index embeddings entirely. We adopt a VGGT-style [58] multi-view transformer with alternating local-global attention, where the local attention boundary naturally defines the association relationship. Different from the original VGGT, E-RayZer performs *pairwise pose prediction*: camera tokens from a canonical view and a target view are concatenated to regress their relative camera pose. Consequently, E-RayZer does

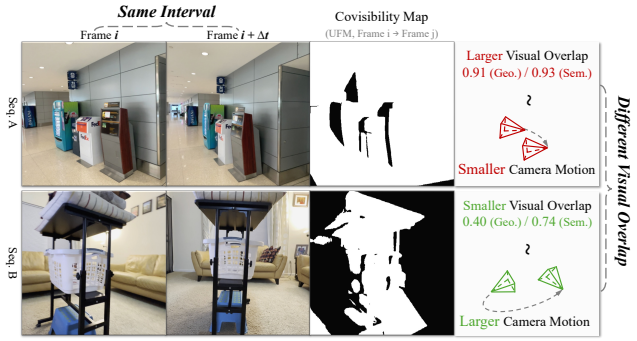

Figure 3. **Different Visual Overlaps under the Same Frame Interval.** Two sequences from DL3DV [34] share the same frame interval yet exhibit drastically different levels of visual overlap. Our proposed semantic and geometric overlap metrics more accurately capture the true difficulty (or camera motion) of each sequence.

not require different camera/register tokens for canonical and non-canonical views. This architectural design is applied to both the transformer used for camera estimation ($f_{\boldsymbol{\theta}}^{\text{cam}}$) and that for scene reconstruction ($f_{\boldsymbol{\psi}'}^{\text{scene}}$).

3.3. Sequence Curriculum Based on Visual Overlap

As E-RayZer leverages explicit scene representation, it suffers from harder convergence when **trained from scratch**. To stabilize training, we propose a learning curriculum based on the concept of *visual overlap* between input views, providing *fine-grained control* over training data difficulty. This curriculum also adaptively aligns the data distributions across diverse data sources, making E-RayZer more scalable to heterogeneous training resources.

We highlight that E-RayZer's learning curriculum fundamentally differs from that of RayZer, which is based on fixed frame-index intervals. As illustrated in Fig. 3, RayZer's interval-based sampling provides only an inaccurate and inflexible approximation of visual overlap, is hard-coded and thus not scalable to heterogeneous resources.

We then describe the two key steps for constructing our learning curriculum: *data labeling* and *sampling*. We introduce two variants of visual-overlap labeling tools: a *geometric* version that computes actual covisibility, and a *semantic* version as an unsupervised approximation of it.

Labeling. For each training sequence u (from any data resource), we compute a spacing profile by *uniformly* sampling a small set of frame triplets for each spacing Δt, as $\mathcal{T}_u(\Delta t) = \{(i,\, i+\Delta t,\, i+2\Delta t)\}$, and averaging the two pairwise overlaps $o(\cdot,\cdot)$ per triplet:

$$o_{\text{tri}}(i, \Delta t) \;=\; \tfrac{1}{2}\Big(o(i,\, i{+}\Delta t) \;+\; o(i{+}\Delta t,\, i{+}2\Delta t)\Big). \quad (8)$$

Averaging $o_{\text{tri}}(i, \Delta t)$ over all sampled triplets yields the per-sequence profile $O_u(\Delta t)$, characterizing how overlap (and consequently difficulty) varies with frame index spacing.

Training-time Sampling. Given curriculum progress $s \in [0, 1]$, we use a visual overlap lower limit of $o(s) = s\, o_{\min} + (1-s)\, o_{\max}$, so that it decreases over training. We then obtain the sequence-specific spacing $\Delta t_u(s)$ by *looking up* the precomputed table $\{(\Delta t_k,\, O_u(\Delta t_k))\}$ and *linearly interpolating* between the nearest entries. Finally, the temporal span of the sampled sequence follows $t = (K_{\text{ref}} - 1)\, \Delta t_u(s)$.

Instantiations. We instantiate o with two alternatives – *geometric overlap* (UFM [79] covisibility, which is trained with 3D annotations) and *semantic overlap* (DINOv2 [38] cosine similarity, which is trained with self-supervision):

$$\begin{aligned} o_{\text{sem}}(i,j) &= \cos\big(\phi_{\text{DINO}}(I_i),\, \phi_{\text{DINO}}(I_j)\big), \\ o_{\text{geo}}(i,j) &= \text{Cov}_{\text{UFM}}(I_i,\, I_j). \end{aligned} \quad (9)$$

In Sec. 4.4, we show that both the semantic and geometric curricula outperform RayZer's interval-based curriculum, and that the two variants perform comparably.

4. Experiments

We first describe the experimental setups in Sec. 4.1. We then evaluate E-RayZer in two aspects: as a self-supervised model for pose estimation and 3D reconstruction (Sec. 4.2), and as a spatial visual pre-training framework for downstream tasks (Sec. 4.3). Finally, we ablate the key design choices of E-RayZer (Sec. 4.4).

4.1. Experimental Setup

Implementation Details. E-RayZer is trained with 10 input images, where 5 are used as reference views and 5 as target views. During training, we follow a linear decay in visual-overlap scores: $1.0 \rightarrow 0.5$ for geometric-overlap scheduling and $1.0 \rightarrow 0.75$ for semantic-overlap scheduling. For a fair comparison, we align RayZer with E-RayZer using the better model architecture and the novel training curriculum. For other baselines, we use official checkpoints and provide specific implementation details in the corresponding subsections. See more details in the supplementary material.

Metrics. For pose estimation, we report relative pose accuracy (RPA) at thresholds of 5°, 15°, and 30°, which jointly reflects rotation and translation accuracy. For novel-view synthesis, we use standard PSNR. For depth estimation, we evaluate absolute relative error (AbsRel) and $\delta < 1.25$, following Depth Anything [68]. For pairwise flow prediction, we report the average end-point error (EPE) and the proportion of outlier flow predictions under thresholds of 1px, 2px, and 5px, following UFM [79].

Datasets. *Training.* We present results of E-RayZer trained on both single-dataset and multi-dataset settings. The single-dataset variants are trained exclusively on RealEstate10K [46] or DL3DV [34], while the multi-dataset variant is trained on a mixture of seven datasets: DL3DV [34], CO3Dv2 [43], RealEstate10K [83], MVImgNet [75], ARKitScenes [6], WildRGB-D [66], and ACID [35], covering diverse indoor and outdoor sequences.

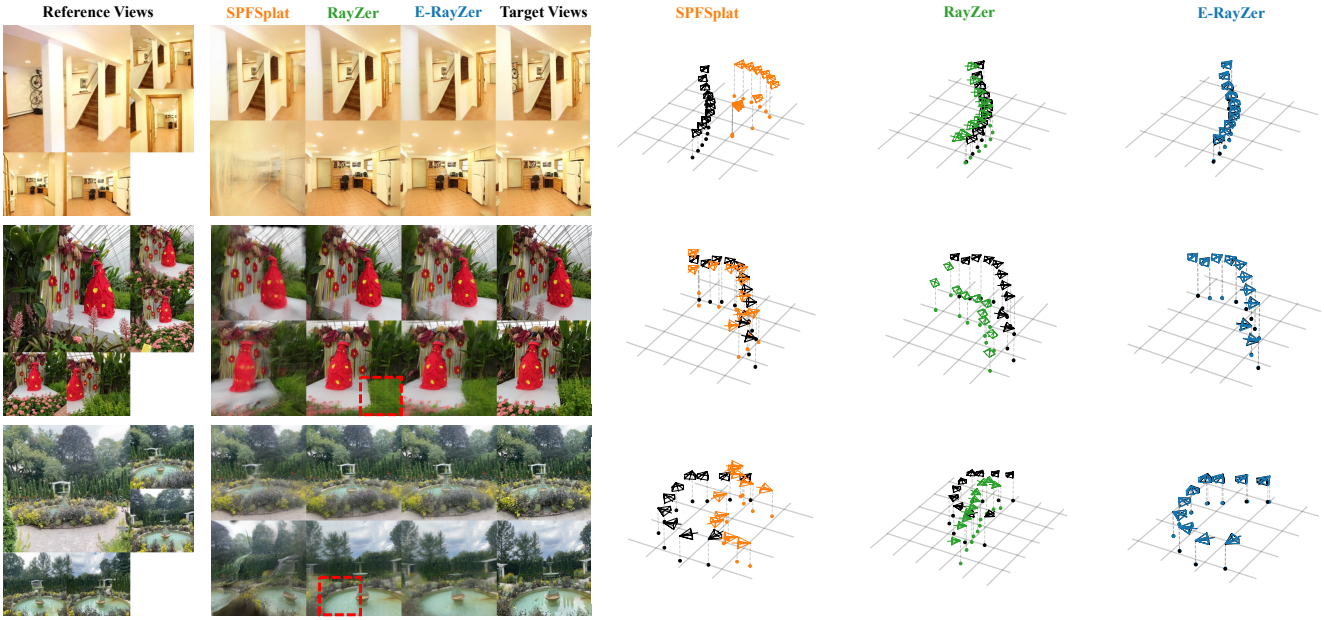

Figure 4. **Visual Comparison with (Partially) Self-supervised Methods.** We include results on both novel-view synthesis (left) and pose estimation (right), where E-RayZer outperforms baselines on pose accuracy, showing its grounded 3D understanding. E-RayZer also outperforms RayZer on low-texture regions (highlighted w/ red box) on NVS, a case where RayZer's view interpolation cannot handle.

Table 1. **Comparison with (Partially) Self-supervised Methods on Novel-view Synthesis (NVS) and Pose Estimation.** We report PSNR for NVS and RPA↑@5°/15°/30° for pose estimation. RayZer [25] and E-RayZer are fully self-supervised methods trained from scratch, while SPFSplat [20] is initialized from MASt3R [14], which itself is trained under dense 3D supervision on 14 datasets.

Method	Self-supervised?	Training Data	WildRGB-D [66]				ScanNet++ [73]				DL3DV [34]			
			PSNR↑	@5°↑	@15°↑	@30°↑	PSNR↑	@5°↑	@15°↑	@30°↑	PSNR↑	@5°↑	@15°↑	@30°↑
SPFSplat [21]	✗ (MASt3R ini.)	RE10K [83] (+ extra)	16.7	31.5	58.0	69.8	14.0	**2.5**	11.8	30.3	15.1	19.5	40.6	50.5
E-RayZer (ours)	✓	RE10K [83]	**21.0**	**40.3**	**89.4**	**96.5**	**17.5**	1.1	**13.3**	**37.3**	**17.3**	**21.2**	**55.0**	**72.7**
RayZer [25]	✓	DL3DV [34]	**25.9**	0.0	0.2	6.5	**20.5**	0.0	0.7	6.2	**21.4**	0.0	0.6	6.2
E-RayZer (ours)	✓		24.3	**84.5**	**98.4**	**99.3**	20.1	**7.7**	**33.6**	**63.0**	20.3	**72.0**	**88.4**	**93.5**
RayZer [25]	✓	7 datasets	**26.7**	0.2	9.3	43.6	**21.5**	0.0	0.9	9.0	**20.8**	0.0	1.9	17.0
E-RayZer (ours)	✓		24.9	**90.8**	**98.6**	**99.3**	20.7	**5.7**	**34.8**	**63.7**	19.7	**59.9**	**82.9**	**90.2**

Evaluation. We primarily evaluate pose estimation and novel-view synthesis on WildRGB-D, DL3DV test set, and the out-of-distribution (OOD) ScanNet++ [73]. To assess the generalization of the learned representations (Sec. 4.3), we evaluate on OOD ScanNet++ and BlendedMVS [69] for pose and depth estimation, and StaticThings3D [48] for pairwise flow prediction.

4.2. Pose Estimation and Novel-view Synthesis

Baselines and Setups. We compare against SPFSplat [21] and RayZer [25]. Notably, SPFSplat is initialized from the supervised MASt3R [32] model, and thus is not truly self-supervised; while E-RayZer and RayZer are trained from scratch under self-supervision. We evaluate pose accuracy on all images and assess novel-view synthesis quality on the target views rendered with predicted camera poses.

Results. As shown in Tab. 1, E-RayZer consistently outperforms SPFSplat [21] on most metrics, despite being truly self-supervised. Moreover, E-RayZer significantly surpasses RayZer [25] in pose estimation across all setups while achieving comparable novel-view synthesis quality. These results suggest that the *explicit* 3D modeling in E-RayZer yields more geometrically meaningful pose representations, whereas RayZer's *implicit* approach overoptimizes for view synthesis quality without being truly 3D-aware, resulting in a less interpretable pose space. Qualitative comparisons in Fig. 4 further support these findings.

4.3. E-RayZer as Self-supervised Pre-training

We validate E-RayZer as a self-supervised spatial visual pre-training framework. First, we show that its performance is comparable to the supervised VGGT and that E-RayZer pre-training further enhances VGGT (Sec. 4.3.1). We then probe the learned features on downstream tasks to verify E-RayZer's representation quality (Sec. 4.3.2).

4.3.1. E-RayZer Initialization Benefits Supervised Model

Baselines and Setups. We compare with the state-of-the-art supervised model VGGT [58]. Note that we train it using the same data and architecture with E-RayZer for an apple-to-apple comparison, denoted as VGGT*.

Table 2. **Comparison with Supervised VGGT [58] on Pose Estimation. E-RayZer's pre-training improves VGGT performance** (last row), forming an effective self-supervised pre-training and supervised post-training paradigm. We report pose accuracy RPA$_\uparrow$@5°/15°. Both models are trained on DL3DV [34] and evaluated on DL3DV & eight out-of-domain datasets for zero-shot testing. Models are labeled as self-supervised or supervised. VGGT* denotes our re-implementation with E-RayZer's pairwise camera head. Results are color-ranked from red to yellow, and we underline the results that our self-supervised E-RayZer surpasses supervised VGGT*.

Method	*In-domain* DL3DV [43]		*Out-of-domain (Zero-shot Generalization)* RE10K [83]		CO3Dv2 [43]		WildRGB-D [66]		7-Scenes [49]		CamLand [29]		BlendedMVS [69]		NAVI [22]		ScanNet++ [73]	
	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$
E-RayZer (ours)	72.0	88.4	83.0	96.8	19.1	61.8	51.1	82.3	38.8	78.0	18.1	62.9	22.9	46.8	20.7	57.8	7.7	33.6
VGGT*	79.6	94.2	80.4	97.9	16.0	64.3	32.5	76.2	34.7	**83.6**	11.1	49.8	17.0	42.8	14.3	54.5	6.7	39.8
E-RayZer→VGGT*	**87.3**	**96.6**	**85.3**	**98.4**	**25.3**	**72.2**	**56.2**	**91.4**	**43.8**	82.8	**30.2**	**75.6**	**29.2**	**52.2**	**26.9**	**64.3**	**14.3**	**53.8**

Table 3. **Probing 3D Spatial Awareness of Learned Representations on Multi-view Depth and Pose Estimation.** We evaluate the learned representations via both frozen-backbone and fully supervised finetuning on ScanNet++ [73] and BlendedMVS [69], which are not included in pre-training for any model. The best results are shown in **bold**, and the second-best are underlined. The experiments only use the *encoders* of RayZer [25] and E-RayZer.

		Method	*Depth* AbsRel$_\downarrow$	*Depth* δ<1.25$_\uparrow$	*Camera Pose* RPA@5°$_\uparrow$	*Camera Pose* RPA@15°$_\uparrow$
ScanNet++ [73]	Frozen	DINOv2 [38]	0.193	74.9	0.8	9.5
		DINOv3 [50]	0.201	73.2	0.4	10.0
		Percep. Encoder [7]	0.203	73.2	0.5	8.5
		CroCo v2 [64]	0.203	73.0	1.4	15.1
		VideoMAE V2 [59]	0.175	76.3	0.1	6.6
		RayZer [25]	0.161	79.3	4.7	27.4
		E-RayZer (ours)	**0.116**	**87.1**	**13.8**	**49.5**
	Full-finetune	DINOv2 [38]	0.178	78.2	3.3	19.6
		DINOv3 [50]	0.176	78.7	4.0	22.3
		Percep. Encoder [7]	0.181	77.8	2.9	20.0
		CroCo v2 [64]	0.177	78.2	3.8	20.8
		VideoMAE V2 [59]	0.076	93.9	12.8	51.4
		RayZer [25]	0.077	93.9	21.5	60.6
		E-RayZer (ours)	**0.059**	**95.1**	**22.7**	**64.3**
BlendedMVS [69]	Frozen	DINOv2 [38]	0.366	50.5	1.1	8.0
		DINOv3 [50]	0.397	49.1	1.2	6.8
		Percep. Encoder [7]	0.385	49.9	1.2	6.2
		CroCo v2 [64]	0.412	47.7	1.6	12.6
		VideoMAE V2 [59]	0.371	49.4	1.0	6.2
		RayZer [25]	0.351	52.6	16.7	34.5
		E-RayZer (ours)	**0.245**	**68.3**	**26.5**	**45.8**
	Full-finetune	DINOv2 [38]	0.353	52.5	1.8	12.8
		DINOv3 [50]	0.349	52.1	1.7	15.3
		Percp. Encoder [7]	0.370	50.3	2.1	11.6
		CroCo v2 [64]	0.369	51.2	2.8	15.9
		VideoMAE V2 [59]	0.197	75.9	17.3	45.5
		RayZer [25]	0.194	77.7	26.1	50.2
		E-RayZer (ours)	**0.148**	**82.8**	**36.2**	**58.8**

Table 4. **Probing 2.5D Spatial Awareness of Learned Representations on Pairwise Flow Estimation.** We evaluate on StaticThings3D [48], an out-of-distribution synthetic dataset. All models are fully finetuned under flow supervision. The best results are shown in **bold**, and the second-best are underlined.

Method	*Error* EPE$_\downarrow$	*Outlier Ratio* @1px$_\downarrow$	@2px$_\downarrow$	@5px$_\downarrow$
CroCo v2 [64]	1.273	17.7	8.7	3.8
VideoMAE V2 [59]	2.028	42.7	22.1	6.9
RayZer [25]	**1.105**	**13.4**	**6.6**	**2.8**
E-RayZer (ours)	1.254	16.9	7.8	3.1

E-RayZer is Comparable with Supervised VGGT*. First two rows of Tab. 2 show that E-RayZer outperforms VGGT* on several out-of-domain datasets (*e.g.*, WildRGB-D [66], CamLand [29], and BlendedMVS [69]). Moreover, E-RayZer almost consistently achieves higher accuracy on RPA@5°, a stricter metric, suggesting better precision in pose prediction. The results demonstrate the strong performance of E-RayZer as a self-supervised method without using any 3D annotations for training.

Effectiveness of Pre-training. As shown in last two rows of Tab. 2, initializing VGGT* with E-RayZer weights yields significant improvements over training from scratch, confirming that E-RayZer serves as an effective pre-training framework for visual geometry learning. The results also suggest that the learned knowledge of our self-supervised and supervised methods are highly complementary (they are trained on same data but pre-training still helps), showing the great potential of spatial visual pre-training.

4.3.2. Probing Representations on Downstream Tasks

Baselines and Setups. To further assess the spatial awareness, we probe and compare the feature representations of E-RayZer against widely-used vision encoders: DINO series [38, 50], CroCo v2 [64], VideoMAE V2 [59], Perception Encoder [7], and RayZer [25]. We only use the backbones and train the prediction heads **from scratch**. We compare performance under both frozen-backbone and full-finetuning settings on downstream tasks, including:

• *Multi-view Depth and Pose Estimation (3D Tasks).* For depth estimation, we apply a DPT head [42] on top of the backbones. Single-view backbones lack multi-view correspondence and thus reduce to monocular depth estimation. For pose estimation, we attach VGGT's [58] camera head to each backbone, using either the class token or averaged patch features as camera tokens. These tokens are aggregated across views via transformer layers, enabling even single-view models to reason over multi-view geometry. We note that the camera heads of RayZer and E-RayZer from their pre-training stage are *not* used for fairness.

• *Pairwise Flow Estimation (2.5D Task).* We consider backbones that encode binocular geometry, including CroCo v2 [64], VideoMAE V2 [59], RayZer [25], and E-RayZer. We follow the settings of UFM [79].

Results on 3D Downstream Tasks. Tab. 3 shows that E-RayZer achieves the best performance across all datasets and settings, demonstrating strong 3D-awareness in its feature representations. Under the frozen-backbone setting, E-

Table 5. **Ablation on Data Mixing and Scaling.** We compare our E-RayZer with supervised VGGT* [58] on varying training data settings. We color-rank the results from red to yellow **for each model itself** across training data, thus **the color distribution reflects their scaling behavior**. We also underline the results where self-supervised E-RayZer outperforms supervised VGGT* (for each training data).

Training Data	Method	NAVI [22]				CO3Dv2 [43]				ScanNet++ [73]				DL3DV [34]			
		PSNR↑	@5°↑	@15°↑	@30°↑	PSNR↑	@5°↑	@15°↑	@30°↑	PSNR↑	@5°↑	@15°↑	@30°↑	PSNR↑	@5°↑	@15°↑	@30°↑
RE10K [83]	VGGT*	/	0.4	8.4	22.5	/	0.1	3.7	15.5	/	0.6	10.0	30.7	/	17.8	50.9	69.4
	E-RayZer	17.2	1.8	16.9	34.0	19.1	0.6	8.3	26.0	17.5	1.1	13.3	37.3	17.3	21.2	55.0	72.7
DL3DV [34]	VGGT*	/	14.3	54.5	75.7	/	16.0	64.3	82.1	/	6.7	39.8	71.5	/	79.6	94.2	97.1
	E-RayZer	20.5	20.7	57.8	69.6	22.9	19.1	61.8	78.8	20.1	7.7	33.6	63.0	20.3	72.0	88.4	93.5
7-dataset Mix	VGGT*	/	28.8	67.3	84.4	/	43.4	83.5	91.8	/	13.1	54.8	78.5	/	66.1	88.9	95.6
	E-RayZer	20.6	24.6	56.1	69.2	24.3	30.3	74.2	83.7	20.7	5.7	34.8	63.7	19.7	59.9	82.9	90.2

RayZer notably outperforms all baselines. With full finetuning, E-RayZer further improves across all metrics, surpassing RayZer [25] and VideoMAE V2 [59] by a large margin. The consistently strong results highlight the generalization ability of its geometrically grounded representations, showing its potential as a spatial visual pre-training framework.

Results on Pairwise Flow Estimation. Tab. 4 shows that E-RayZer achieves competitive performance on pairwise flow prediction, closely following RayZer [25], despite not being trained directly for tasks that optimize image correspondences (*e.g.*, masked image modeling in CroCo v2 [64] and VideoMAE V2 [59], or view interpolation in RayZer). Compared to E-RayZer, RayZer holds a slight advantage due to its implicit 3D formulation, naturally suited for low-level motion estimation. Nevertheless, E-RayZer outperforms other baselines, demonstrating that its explicit 3D representation learning captures meaningful spatial correspondences even for the 2.5D task.

4.4. Ablation Study

Data Mixing / Scaling. We investigate the behavior of self-supervised E-RayZer and supervised VGGT* (Sec. 4.3.1) under varying data scales and quality. In Tab. 5, E-RayZer and VGGT* demonstrate a similar scaling behavior: training on data with broader distributions improves generalization (*e.g.*, models trained on 7 datasets outperform those trained on DL3DV alone). However, reducing the sampling frequency of a particular domain slightly degrades performance on its corresponding test set (*e.g.*, 7-dataset models perform worse on DL3DV than DL3DV-only models), a trend consistently observed in prior work [16, 67, 71]. Besides, data quality also plays a key role, as training on DL3DV yields better results than that on RE10K.

Moreover, again, the self-supervised model (E-RayZer) achieves performance on par with the supervised VGGT* (while VGGT* holds advantage when trained on large data), demonstrating that large-scale self-supervision alone can yield geometrically grounded 3D understanding. This result underscores that data diversity and quality, rather than explicit 3D supervision, are the true drivers of scalability in large 3D vision models. Together, these results highlight the great potential of self-supervised 3D learning when scaled to internet-scale data, and provide valuable guidance for future data selection and curation strategies.

Table 6. **Ablation on Curriculum Learning.** We compare four curriculum strategies when training E-RayZer on DL3DV (top) and a seven-dataset mixture (bottom). The proposed visual-overlap-based curriculum consistently outperforms baselines.

	Curriculum Variant	PSNR↑	RPA@5°↑	RPA@15°↑	RPA@30°↑
DL3DV	No Curriculum	16.1	4.0	27.8	47.2
	Frame Interval	19.8	56.1	79.3	86.0
	Semantic Overlap	**20.4**	**73.2**	**88.7**	**93.7**
	Geometric Overlap	20.3	72.0	88.4	93.5
7-dataset	No Curriculum	15.9	2.1	21.6	40.7
	Frame Interval	19.1	43.8	72.1	82.9
	Semantic Overlap	**19.7**	58.7	81.0	89.8
	Geometric Overlap	**19.7**	**59.9**	**82.9**	**90.2**

Curriculum Learning. In Tab. 6, we compare against two baselines with (1) no curriculum, and (2) a frame-interval-based curriculum, where frame intervals are specified for each dataset. Across two training regimes (*i.e.*, DL3DV-only and the seven-dataset mixture), the proposed visual-overlap curricula consistently outperform both baselines, with the two variants performing comparably. These results demonstrate that our fine-grained curriculum strategy significantly improves self-supervised pose estimation and reconstruction, while eliminating the need for manual tuning for each training dataset and benefiting scaling.

5. Discussion

We propose E-RayZer, a multi-view 3D model that learns geometrically grounded representations via self-supervised 3D reconstruction. While our results are promising, we identify several limitations and directions for future work.

E-RayZer currently operates on static scenes, limiting its training to existing static datasets – high-quality static-scene data remains scarce beyond these benchmarks. Extending it to dynamic scenes would enable learning from generic videos and is a promising direction. Additionally, our learning curriculum generally assumes continuous video frames with fairly uniform camera motion; sparse images or frames with drastic viewpoint changes may reduce its effectiveness.

Despite these limitations, E-RayZer outperforms prior self-supervised methods and is competitive with supervised approaches. Extensive experiments show that E-RayZer pre-training benefits supervised models and other 3D downstream tasks, establishing it as a scalable 3D-aware visual pre-training framework.

Acknowledgements. The work is partially done during Qitao Zhao's internship at Adobe Research. We thank Zhengqi Li for insightful advice. We also thank Frédéric Fortier-Chouinard, Jiashun Wang, Yanbo Xu, Zihan Wang, and members of the Physical Perception Lab for helpful discussions.

This work was also supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 3

[2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 2

[3] Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *ECCV*, 2018. 2

[4] Mohamed El Banani, Jason J Corso, and David F Fouhey. Novel object viewpoint estimation through reconstruction alignment. In *CVPR*, 2020. 2

[5] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. 3

[6] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *NeurIPS D&B*, 2021. 5

[7] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. In *NeurIPS*, 2025. 1, 2, 7

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 2

[9] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *CVPR*, 2021. 2

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[12] Yu Chen and Gim Hee Lee. Dbarf: Deep bundle-adjusting generalizable neural radiance fields. In *CVPR*, 2023. 3

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2

[14] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *3DV*, 2025. 2, 6

[15] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022. 3

[16] Negar Foroutan, Paul Teiletche, Ayush Kumar Tarun, and Antoine Bosselut. Revisiting multilingual data mixtures in language model pretraining. *arXiv preprint arXiv:2510.25947*, 2025. 8

[17] Yang Fu, Ishan Misra, and Xiaolong Wang. Mononerf: Learning generalizable nerfs from monocular videos without camera poses. In *ICML*, 2023. 3

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2, 3

[20] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. Pf3plat: Pose-free feed-forward 3d gaussian splatting for novel view synthesis. In *ICML*, 2025. 2, 6

[21] Ranran Huang and Krystian Mikolajczyk. No pose at all: Self-supervised pose-free 3d gaussian splatting from sparse views. In *ICCV*, 2025. 2, 6

[22] 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. In *NeurIPS*, 2023. 7, 8

[23] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. In *3DV*, 2024. 2

[24] Hanwen Jiang, Zhenyu Jiang, Yue Zhao, and Qixing Huang. LEAP: Liberate sparse-view 3d modeling from camera poses. In *ICLR*, 2024. 2

[25] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. In *ICCV*, 2025. 1, 2, 3, 4, 6, 7, 8

[26] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. In *ACM SIGGRAPH Asia*, 2025. 2

[27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 3

[28] Gyeongjin Kang, Jisang Yoo, Jihyeon Park, Seungtae Nam, Hyeonsoo Im, Sangheon Shin, Sangpil Kim, and Eunbyung Park. Selfsplat: Pose-free and 3d prior-free generalizable 3d gaussian splatting. In *CVPR*, 2025. 2

[29] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *CVPR*, 2017. 7

[30] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM ToG*, 2023. 2, 3, 4

[31] Zihang Lai, Sifei Liu, Alexei A Efros, and Xiaolong Wang. Video autoencoder: self-supervised disentanglement of static 3d structure and motion. In *ICCV*, 2021. 3

[32] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 6

[33] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. In *3DV*, 2024. 2

[34] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024. 5, 6, 7, 8

[35] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *ICCV*, 2021. 5

[36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[37] Thomas W Mitchel, Hyunwoo Ryu, and Vincent Sitzmann. True self-supervised novel view synthesis is transferable. In *ICLR*, 2026. 3

[38] 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. In *TMLR*, 2024. 2, 3, 5, 7

[39] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 3

[40] Julius Plucker. Xvii. on a new geometry of space. In *Philosophical Transactions of the Royal Society of London*, 1865. 3

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[42] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 7

[43] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 5, 7, 8

[44] Chris Rockwell, Justin Johnson, and David F Fouhey. The 8-point algorithm as an inductive bias for relative pose prediction by vits. In *3DV*, 2022. 2

[45] Mehdi SM Sajjadi, Aravindh Mahendran, Thomas Kipf, Etienne Pot, Daniel Duckworth, Mario Lučić, and Klaus Greff. Rust: Latent neural scene representations from unposed imagery. In *CVPR*, 2023. 3

[46] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. In *CVPR*, 2024. 5

[47] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2

[48] Philipp Schröppel, Jan Bechtold, Artemij Amiranashvili, and Thomas Brox. A benchmark and a baseline for robust multi-view depth estimation. In *3DV*, 2022. 6, 7

[49] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *CVPR*, 2013. 7

[50] 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. 1, 2, 3, 7

[51] Samarth Sinha, Jason Y Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B Lindell. Sparsepose: Sparse-view camera pose regression and refinement. In *CVPR*, 2023. 2

[52] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs, 2024. arXiv preprint arXiv:2408.13912. 2

[53] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 2, 3

[54] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. In *NeurIPS*, 2023. 3

[55] Khiem Vuong, Anurag Ghosh, Deva Ramanan, Srinivasa Narasimhan, and Shubham Tulsiani. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In *CVPR*, 2025. 2

[56] Haoru Wang, Kai Ye, Yangyan Li, Wenzheng Chen, and Baoquan Chen. The less you depend, the more you learn: Synthesizing novel views from sparse, unposed images without any 3d knowledge. In *ICLR*, 2026. 3

[57] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *ICCV*, 2023. 2

[58] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 1, 2, 4, 6, 7, 8

[59] 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 *CVPR*, 2023. 1, 2, 7, 8

[60] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 2

[61] Ruoyu Wang, Yi Ma, and Shenghua Gao. Recollection from pensieve: Novel view synthesis via learning from uncalibrated videos. *arXiv preprint arXiv:2505.13440*, 2025. 3

[62] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2

[63] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. In *NeurIPS*, 2022. 3

[64] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *ICCV*, 2023. 1, 2, 3, 7, 8

[65] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, 2020. 3

[66] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *CVPR*, 2024. 5, 6, 7

[67] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *NeurIPS*, 2023. 8

[68] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 5

[69] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020. 6, 7

[70] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *ICLR*, 2025. 2

[71] Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *ICLR*, 2025. 8

[72] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, et al. gsplat: An open-source library for gaussian splatting. In *JMLR*, 2025. 4

[73] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 6, 7, 8

[74] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 2

[75] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023. 5

[76] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *ECCV*, 2022. 2

[77] Jason Y. Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Sparse-view pose estimation via ray diffusion. In *ICLR*, 2024. 2, 3

[78] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *ECCV*, 2024. 2

[79] Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, et al. Ufm: A simple path towards unified dense correspondence with flow. In *NeurIPS*, 2025. 5, 7

[80] Qitao Zhao and Shubham Tulsiani. Sparse-view pose estimation and reconstruction via analysis by generative synthesis. In *NeurIPS*, 2024. 2

[81] Qitao Zhao, Amy Lin, Jeff Tan, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Diffusionsfm: Predicting structure and motion via ray origin and endpoint diffusion. In *CVPR*, 2025. 2

[82] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 3

[83] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *ACM SIGGRAPH*, 2018. 5, 6, 7, 8

[84] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 2