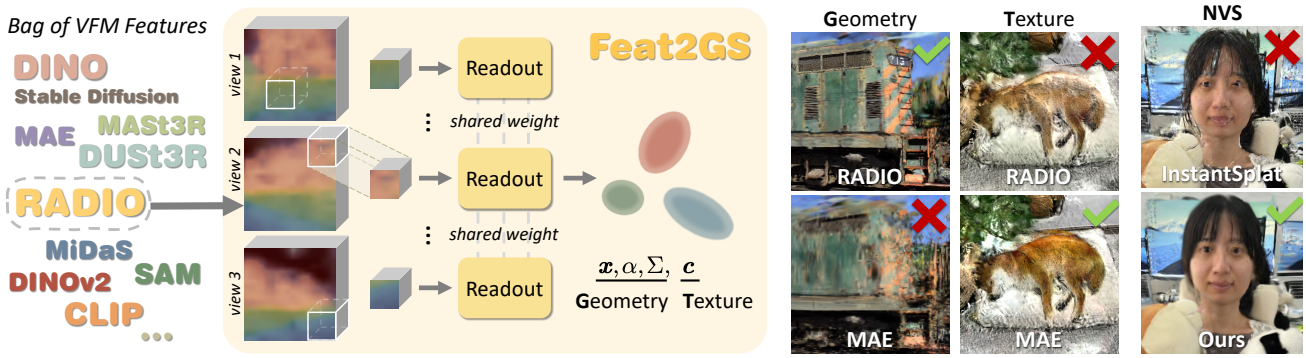
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Feat2GS: Probing Visual Foundation Models with Gaussian Splatting

Yue Chen[1,2] Xingyu Chen[1,2] Anpei Chen[2,3] Gerard Pons-Moll[3,4] Yuliang Xiu[2,5]

[1]Zhejiang University [2]Westlake University [3]University of Tübingen, Tübingen AI Center
[4]Max Planck Institute for Informatics, Saarland Informatics Campus
[5]Max Planck Institute for Intelligent Systems

fanegg.github.io/Feat2GS



(a) Feat2GS as VFM Probe (b) Extensive Analysis of VFM (c) Strong Baseline

Figure 1. We present **Feat2GS**, a unified framework to probe "texture and geometry awareness" of visual foundation models (VFMs). Novel-view synthesis (NVS) serves as an effective proxy for 3D evaluation. (a) Casually captured photos are input into VFMs to extract features and into a stereo reconstructor to obtain relative poses. Pixel-wise features are transformed into 3D Gaussians (3DGS) using a lightweight readout layer trained with photometric loss. (b) 3DGS parameters, grouped into **G**eometry and **T**exture, enable separate analysis of geometry/texture awareness in VFMs, evaluated by the NVS quality on diverse, unposed open-world images. (c) Our baseline derived from extensive empirical analysis, achieves superior performance for NVS by simply concatenating features from diverse VFMs.

Abstract

Given that visual foundation models (VFMs) are trained on extensive datasets but often limited to 2D images, a natural question arises: how well do they understand the 3D world? With the differences in architecture and training protocols (i.e., *objectives, proxy tasks), a unified framework to fairly and comprehensively probe their 3D awareness is urgently needed. Existing works on 3D probing suggest single-view 2.5D estimation* (e.g., *depth and normal) or two-view sparse 2D correspondence* (e.g., *matching and tracking). Unfortunately, these tasks ignore texture awareness, and require 3D data as ground-truth, which limits the scale and diversity of their evaluation set. To address these issues, we introduce Feat2GS, which readout 3D Gaussians attributes from VFM features extracted from unposed images. This allows us to probe 3D awareness for geometry and texture via novel view synthesis, without requiring 3D data. Additionally, the disentanglement of 3DGS parameters – geometry ($\boldsymbol{x}, \alpha, \Sigma$) and texture ($\boldsymbol{c}$) – enables separate analysis of texture and geometry awareness. Under Feat2GS, we conduct extensive experiments to probe the 3D awareness of several VFMs, and investigate the ingredients that lead to a 3D aware VFM. Building on these findings, we develop several variants that achieve state-of-the-art across diverse datasets. This makes Feat2GS useful for probing VFMs, and as a simple-yet-effective baseline for novel-view synthesis. Code and data will be made available at* fanegg.github.io/Feat2GS.

1. Introduction

Visual foundation models (VFMs) [6] have emerged as the basis for various 2D reasoning tasks [45, 70] and as a critical component for 3D fine-tuning [3, 10, 13, 35, 40, 71, 95, 107, 113]. Their strong few-shot or zero-shot generalization ability stems mainly from the expressive features [9, 73, 76]. But what is the key for the 3D expressiveness? *Does 3D awareness have to come from 3D data?* Some VFMs, such as DINOv2 and MAE, are trained using only

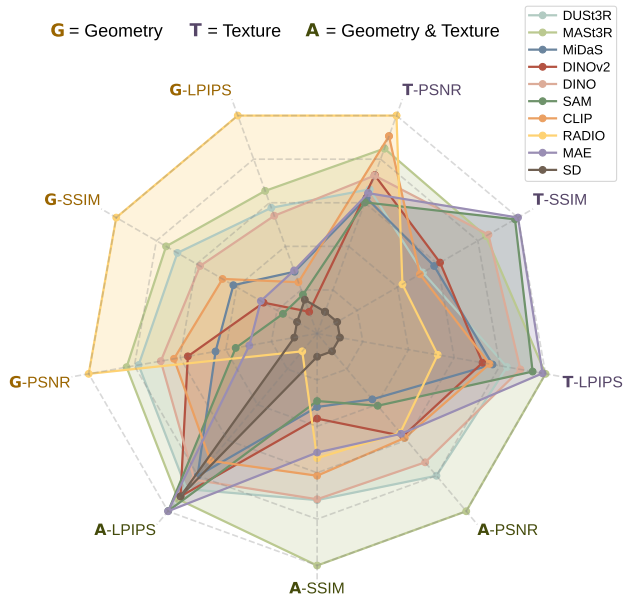


Figure 2. **Texture+Geometry probing of mainstream VFMs.** Normalized average metrics for novel view synthesis (NVS) across six datasets are plotted on axes, with higher values away from the center indicating better performance. Try the interactive visualization demo on fanegg.github.io/Feat2GS.

2D images. *How important is the training approach?* VFMs differ in many aspects, such as learning strategies (*e.g.*, self-supervised [9, 34], supervised learning [45, 50, 95]), and proxy tasks (*e.g.*, depthmap regression [95], cross-view completion [98], generation [76]). These differences make fair and comprehensive benchmarking difficult.

To answer these questions, recent works [22, 75] evaluate the geometry awareness of VFMs using two proxy tasks: 1) 2.5D depth/normal/token estimation from a single image [22, 75], and 2) 2D matching/tracking between two views [2, 22]. Though it does analyze the 3D awareness of current VFMs, it does not probe the texture awareness and multi-view dense consistency of VFM features, which are critical for 3D-related tasks, such as reconstruction and generation.

For **"texture awareness"**, texture-invariant training improves geometry estimation but can harm texture preservation (see RADIO in Fig. 2). However, recovering original textures from VFM features is key for training on large-scale 2D data with photometric loss [24, 35, 82, 106] and texture synthesis [21]. **"Multi-view dense probing"**, like novel view synthesis (NVS) [8, 83], allows every input pixel to contribute to the evaluation, rather than just sparse matching points. Unlike 2D sparse matching, NVS only requires images, eliminating the need for costly labeling of visual correspondences. With the numerous public multi-view datasets [4, 47, 53, 61, 112] available, covering diverse scenes and viewpoints, a new 3D probing approach using these datasets to evaluate texture and geometry awareness in dense mode could be invaluable.

Feat2GS	-**G**eometry	-**T**exture	-**A**ll	InstantSplat [23]
Feature-Readout	$\boldsymbol{x}, \alpha, \Sigma$	$\boldsymbol{c}$	$\boldsymbol{x}, \boldsymbol{c}, \alpha, \Sigma$	-
Free-Optimize	$\boldsymbol{c}$	$\boldsymbol{x}, \alpha, \Sigma$	-	$\boldsymbol{x}, \boldsymbol{c}, \alpha, \Sigma$

Table 1. **GTA probing schemes**. Unlike InstantSplat, Feat2GS uses shallow readout layer to parse VFM features into 3DGS. GTA modes include: **G**eometry– VFM features to Gaussian geometric parameters (*i.e.*, positions $\boldsymbol{x}$, opacity α, covariance Σ), **T**exture– VFM features to Gaussian textural parameters, *i.e.*, SH coefficients $\boldsymbol{c}$, and **A**ll– all parameters are regressed from features.

Thus, we introduce Feat2GS, short for Feature2Gaussian, which evaluates both texture and geometry awareness of VFMs, in the NVS task, using only 2D multi-view data. As shown in Fig. 1, during training, Feat2GS extracts image features from the input views using pre-trained VFMs. A shallow MLP readout layer then regresses the parameters of 3D Gaussians [41] from these features. Multi-view photometric loss minimizes the visual difference between renderings and inputs. During testing, visual similarity metrics (*i.e.*, PSNR, SSIM, LPIPS) are measured for unseen views, across diverse datasets, with Tab. 4 demonstrating that these 2D metrics align well with 3D metrics. To handle sparse and uncalibrated casual images, we initialize camera parameters using DUSt3R [95] and refine them with photometric loss.

The parameters of 3DGS, grouped into geometry ($\boldsymbol{x}, \alpha, \Sigma$) and texture ($\boldsymbol{c}$), enable separate analysis of VFM's texture and geometry awareness. Each group could switch between the "Feature-Readout" and "Free-Optimize" modes to use VFM features as input or free-optimized. This leads to three probing schemes (short as GTA), as shown in Tab. 1.

In summary, our key contributions are as follows:
1) Feat2GS as VFM probe. Feat2GS offers a unified framework (Fig. 1) to probe the 3D awareness (texture and geometry) of pre-trained VFMs, without using 3D labels.
2) Extensive analysis of VFM. We evaluate a wide range of mainstream VFMs (Fig. 2) across diverse multi-view datasets (Tab. 5), spanning from simple scenes to causal captures. These experiments reveal common drawbacks of VFMs and shed light on how to improve them (Sec. 4.3).
3) Strong baseline for NVS. Motivated by these findings, we design three variants of Feat2GS that outperform the current SOTA InstantSplat [23] in all metrics (Tab. 6).

2. Related Work

Measuring 3D Awareness of VFMs. There is no doubt that, visual foundation models [6], short as VFMs, have significantly advanced various 3D vision tasks, such as geometric cue estimation [3, 28, 36, 40, 43, 71, 107], 6D pose estimation [66], visual tracking [89], and spatial reasoning [13, 25], *etc*. However, behind these advances and everyday SOTA records, *are these VFMs truly 3D-aware, even when trained without any 3D data? If so, to what extent? And what enables such awareness?* There is a line of works that try to

answer these questions through multi-view object consistency [7], spatial visual question answering [27, 122], visual perspective taking [54] and robot learning [57, 79, 119]. Although in the same spirit of 3D probing, these existing works mainly focus on coarse-grained semantic reasoning, such as determining "which marker is closer" instead of fine-grained, or even pixel-wise spatial reasoning, like depth estimation. Regarding the fine-grained 3D probes, they either use 2.5D proxy tasks, such as geometric cues estimation (*i.e.*, depth, normal) [22] and view token estimation [75], or use two-view sparse point matching [22] and tracking [2] to assess the 3D awareness of VFMs. The main constraint of these fine-grained 3D probes is their reliance on labeled 3D data, which significantly limits fair and comprehensive evaluation on large-scale visual data. Feat2GS addresses this by first regressing 3DGS from VFM features and then benchmarking 3D awareness via novel view synthesis. This comes with two advantages: ALL raw pixels can contribute to the final evaluation, and ANY multi-view captures can be leveraged. Feat2GS enables "dense" and "diverse" 3D probing.

NVS from Casual Images. Novel view synthesis has made significant progress in recent years [11, 41, 62, 63, 86]. When it comes to sparse and causal captures, which is a quite challenging scenerio, various regularizers [64, 103] or visual priors, such as depth [20, 120], pre-trained visual features [37, 102], diffusion priors [97, 99] and feed-forward modeling [10, 12, 17, 35, 39, 55, 77, 90, 111, 121], have been introduced. However, these methods assume known camera poses from Structure-from-Motion [78], which are not available for sparse captures with minimal overlap. Although some works attemp to optimize camera poses alongside NVS optimization [96], using techniques like coarse-to-fine encoding [52], local-to-global registration [16], geometric constrains [38], adversarial objectives [59], dense correspondence [87], and external priors [5, 29, 60], they can only handle dense-view or video-like sequences — not sparse-view images. Groundbreaking methods like DUSt3R [95], MASt3R [50], and subsequent works [92, 115] address these limitations by training models on large-scale datasets. They approach the pairwise reconstruction problem as a regression of point maps, easing the strict constraints of traditional projective camera models. This enables "Unconstrained Stereo 3D Reconstruction" of arbitrary image collections, without needing prior information about camera calibration or viewpoint poses. The predicted pointmap can directly initialize 3DGS [41], which can then be regressed in a two-view feedforward [24, 82, 106] or optimized with multi-view photometric losses [23]. InstantSplat [23] closely mirrors our target of optimizing 3DGS from sparse captures using DUSt3R estimated cameras. What sets our Feat2GS apart is that we readout 3DGS using visual features, instead of optimizing it in free form, see Tab. 1. This can be done with a shallow readout MLP, helping to prevent overfitting.

3. Method

3.1. Feat2GS

We illustrate our pipeline in Fig. 1. After extracting frozen feature maps from various visual foundation models (VFMs), we take the following steps to ensure fair probing: unifying the feature channel dimensions using Principal Component Analysis (PCA) [33], standardizing the spatial dimensions via bilinear upsampling, and maintaining a consistent network architecture for different VFM features. Specifically, Feat2GS takes the compact features $\mathbf{f}_i$ of each pixel $i \in \{1, 2, \ldots, n\}$ as input and output per-pixel Gaussian primitive via a readout layer g_Θ:

$$\boldsymbol{G}_i = g_\Theta(\mathbf{f}_i) \tag{1}$$

where each Gaussian $\boldsymbol{G}_i$ is parameterized by: position $\boldsymbol{x} \in \mathbb{R}^3$, opacity $\alpha \in \mathbb{R}$, covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$, and three order of spherical harmonic (SH) coefficients $\{\boldsymbol{c}_i \in \mathbb{R}^{48} | i = 1, 2, ..., n\}$.

To ensure the readout layer acts purely as an information conduit rather than a memory storage, we minimize its parameters that forcing the 3D Gaussians are decoded from the features. Specifically, the readout layer is constructed using a 2-layer MLP with 256 units per-layer and ReLU activation instead alternatives like dense prediction transformer [22]. With ReLU activation, it forms the minimal setup for nonlinear mapping. Then we splat 3D Gaussians onto images via differentiable rasterization. Note that, to enable our method to evaluate casually captured, sparse, and uncalibrated images, we use an unconstrained stereo reconstructor [50, 92, 115], DUSt3R [95] in our experiments, to initialize camera poses $\boldsymbol{T}$, which are then jointly updated with the readout layer $g_\Theta^{(mode)}$ or freely-optimized Gaussian parameters $\boldsymbol{O}^{(mode)}$ in a specific mode, using a simple photometric loss between renderings $\mathcal{R}_v(\cdot)$ and images $\{\mathcal{I}_v\}_{v=1}^N$:

$$\Theta^*, \boldsymbol{O}^*, \boldsymbol{T}^* = \underset{\Theta, \boldsymbol{O}, \boldsymbol{T}}{\arg\min} \sum_{v \in N} \| \mathcal{R}_v(g_\Theta(\mathbf{f}), \boldsymbol{O}, \boldsymbol{T}) - \mathcal{I}_v \| \tag{2}$$

To decouple the geometry and texture awareness, we propose three probing modes: **G**eometry reads out geometric parameters from the 2D image features, and freely optimizes textural parameters $\boldsymbol{c}_i$:

$$\{\boldsymbol{x}_i, \alpha_i, \Sigma_i\} = g_\Theta^{(G)}(\mathbf{f}_i), \{\boldsymbol{c}_i\} = \boldsymbol{O}^{(G)} \tag{3}$$

Texture reads out textural parameters, and directly optimizes geometric parameters $\{\boldsymbol{x}_i, \alpha_i, \Sigma_i\}$:

$$\{\boldsymbol{c}_i\} = g_\Theta^{(T)}(\mathbf{f}_i), \{\boldsymbol{x}_i, \alpha_i, \Sigma_i\} = \boldsymbol{O}^{(T)} \tag{4}$$

All reads out all Gaussian parameters:

$$\{\boldsymbol{x}_i, \alpha_i, \Sigma_i, \boldsymbol{c}_i\} = g_\Theta^{(A)}(\mathbf{f}_i), \{\} = \boldsymbol{O}^{(A)} \tag{5}$$

VFM	Arch.	Channel	Supervision	Dataset
DUSt3R [95]	ViT-L/16	1024	Point Regression	3D DUSt3R-Mix
MASt3R [50]	ViT-L/16	1024	Point Regression	3D MASt3R-Mix
MiDaS [71]	ViT-L/16	1024	Depth Regression	3D MiDaS-Mix
DINOv2 [65]	ViT-B/14	768	Self Distillation	2D LVD-142M
DINO [9]	ViT-B/16	768	Self Distillation	2D ImageNet-1k
SAM [45]	ViT-B/16	768	Segmentation	2D SA-1B
CLIP [70]	ViT-B/16	512	Contrastive VLM	2D WIT-400M
RADIO [73]	ViT-H/16	1280	Multi-teacher Distillation	2D DataComp-1B
MAE [34]	ViT-B/16	768	Image Reconstruction	2D ImageNet-1k
SD [76]	UNet	1280	Denoising VLM	2D LAION

Table 2. **VFMs for Evaluation.** For fair comparison, we use checkpoints with comparable architectures and training scales, unify the feature channel dimensions via PCA, and maintain a consistent probing network architecture for all VFMs.

3.2. Warm Start

We find that directly decoding 3D structures from 2D image features can easily stuck in local minimal due to the sparse nature of casual images. To ensure robust evaluation of features from diverse foundation models, we warm up our optimization using a point cloud regression:

$$\min_{\Theta} \|g_{\Theta}(\mathbf{f}) - \boldsymbol{G}_{init}\| \tag{6}$$

where $\boldsymbol{G}_{init}$ refers to the initialization point cloud comes from DUSt3R [95].

3.3. Evaluation

We choose to evaluate on NVS from casual (sparse and uncalibrated) images [23] for two main reasons: (1) Diversity. The capability of handling casual images helps diversify the evaluation data by lowering the requirements for acquisition techniques and view setups. (2) Discrepancy. This task poses more of a challenge compared to dense-view NVS, making it better to differentiate the performance of various VFM features. To enable our evaluation to cover arbitrary casual capturing from 3 to N views, we uniformly estimate the camera parameters of both training and test views across all datasets via unconstrained stereo reconstructor. Subsequently, we perform test-time pose optimization [5, 23, 29, 52, 96] via photometric loss to further refine the test poses before evaluating view synthesis quality.

4. Experiments

4.1. Experimental Setup

Features. We focus our experiments on 10 VFMs that show strong potential for generalizable 3D awareness, comparing models trained on different data types (2D *vs.* 3D) and supervision strategies (*e.g.*, supervised *vs.* self-supervised, point *vs.* depth). An overview is provided in Tab. 2, with more details in *Sup.Mat.* To make the comparison as fair as possible, we use publicly available checkpoints and select those with comparable architectures and training scales. We also incorporate IUVRGB, comprising image index (I), pixel coordinates (UV), and colors (RGB), as a baseline.

Dataset	Scene Type	Complexity	View Range	Views
LLFF [61]	Indoor	Simple	Small	2
DTU [1]	Indoor Object	Simple	Small	3
DL3DV [53]	Indoor / Outdoor	Moderate	Medium	5-6
Casual	Daily Scenario	Moderate	Medium	4-7
MipNeRF360 [4]	Unbounded	Moderate	360	6
MVimgNet [112]	Outdoor Object	Moderate	180-360	2-4
T&T [47]	Indoor / Outdoor	High	Large	6

Table 3. **Datasets for Evaluation.** Classified by scene type, complexity, viewpoint variation, and sampled views.

Datasets. To reliably evaluate different features, our experiments utilize seven multi-view datasets, with sparse views sampled spanning from 2 to 7, and test viewpoints far from the training viewpoints. These datasets, as shown in Tab. 3, rich in diversity, provide us with a more comprehensive perspective compared to datasets with 3D ground-truth.

Metrics. We evaluate novel view synthesis across seven datasets using standard metrics: PSNR, SSIM, and LPIPS. For metric calculation, we follow Splatt3R [82] by applying masks to both the rendered and test images. These masks define valid pixels as those inside the frustum of at least one view and with reprojected depths aligned with DUST3R predicted depth. All metrics are computed over the entire image. On the DTU dataset, we measure the distance between reconstructed 3DGS and point cloud ground truth (Tab. 4), reporting average accuracy, completeness, and distance, as in prior works [1, 92]. Accuracy is the smallest Euclidean distance from a reconstructed point to ground truth, and completeness is the smallest Euclidean distance from a ground-truth point to the reconstruction. Distance is the Euclidean distance based on ground-truth point matching.

4.2. Motivation Validation

Novel View Synthesis Correlates with 3D Metrics. Using 2D metrics instead of 3D ones allows us to bypass the need for 3D ground-truth. Motivated by this, we propose Feat2GS to assess the 3D awareness of VFMs through the 2D metric of novel view synthesis (NVS). The key question is: *Can novel view synthesis effectively serve as a proxy for 3D metrics?* We posit that high-quality NVS strongly correlates with an accurate 3D representation. To validate this hypothesis, we conduct experiments on the DTU dataset [1] with dense pointclouds as 3D ground-truth, evaluating both the 2D NVS and the 3D point cloud regression tasks. We then calculated the correlation between these results, as shown in Tab. 4. The results reveal a strong correlation between 2D and 3D metrics, supporting NVS as an indicator for 3D assessment. We further qualitatively demonstrate this correlation in Fig. 3 (See more details in *Sup.Mat.*'s Fig. R.5). The results indicate a strong relationship between NVS and 3D metrics, confirming that high-quality NVS aligns closely with accurate 3D representations.

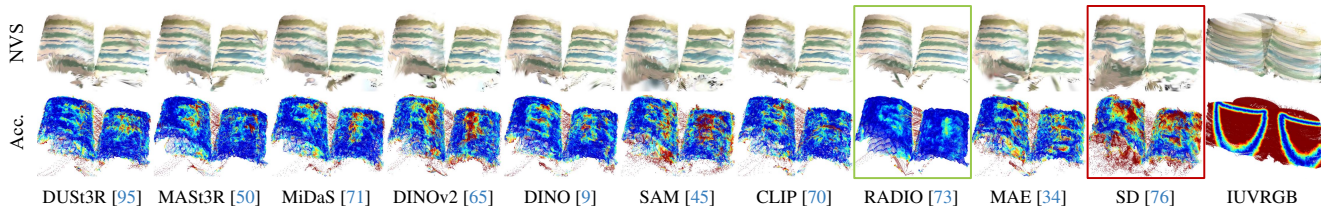
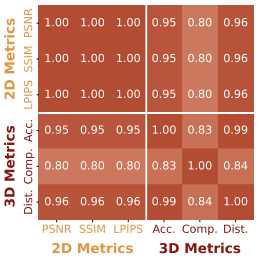


Figure 3. **Novel View Synthesis as Proxy Task to Assess 3D.** We present qualitative examples from the DTU dataset, including NVS and Accuracy. Results show that NVS quality aligns with 3D metrics, proving its reliability as an indicator for 3D assessment. RADIO performs best, SD worst, with IUVRGB as a reference. **Zoom in** or check our webpage to see more details.

	2D Metrics			3D Metrics		
Feature	PSNR↑	SSIM↑	LPIPS↓	Acc.↓	Comp.↓	Dist.↓
DUSt3R	21.36	.7772	.2195	2.439	1.316	6.955
MASt3R	21.44	.7792	.2177	2.321	1.286	6.557
MiDaS	21.09	.7712	.2254	2.934	1.412	8.230
DINOv2	21.01	.7695	.2277	3.101	1.337	8.588
DINO	21.40	.7783	.2187	2.440	1.316	6.885
SAM	20.93	.7660	.2304	3.176	1.339	8.785
CLIP	21.26	.7752	.2215	2.357	1.209	6.739
RADIO	21.78	.7871	.2042	1.886	1.326	5.431
MAE	20.96	.7666	.2289	2.963	1.337	8.374
SD	20.76	.7638	.2343	4.334	1.603	11.594
IUVRGB	16.09	.6825	.3134	13.015	16.957	46.671

(a) 2D Metrics *vs.* 3D Metrics



(b) Correlation Matrix

Table 4. **Novel View Synthesis Aligns Well with 3D Metrics.** (a) We report NVS quality and the Euclidean distance between reconstructed 3DGS positions and pointcloud ground truth on the DTU dataset. (b) Strong 2D-3D metric correlation supports NVS as a benchmark for 3D assessment.

Data Diversity Matters for Comprehensive Probing. Testing on small-scale data can lead to biased conclusions. As shown in Tab. 5, the evaluation results vary across probing GTA modes and different datasets. For instance, LLFF is relatively simple for novel view synthesis due to its dense view capturing and small scale. MASt3R, DUSt3R, and DINO show superior geometry results on LLFF. However, none of them ranks higher than RADIO on T&T dataset, which features more challenging scenes. We sort the datasets from easy to hard in Tab. 5 to derive insights from the variations. Additionally, we show performance correlations in *Sup.Mat.*'s Fig. R.1. We observe that no method consistently performs well across all datasets when the evaluation set is large. Dataset evaluation bias is inevitable. By removing the need for 3D ground truth, we can evaluate on large-scale diverse captures, thereby ensuring that the results are much less biased. Therefore, we base our findings on overall performance and discuss outliers separately.

4.3. Findings

Overall Performance. Table 5 benchmarks VFM features with three probing modes: **G**eometry, **T**exture and **All**. The mean scores across diverse datasets are plotted in Fig. 2. The top three performers in **G**eometry mode are RADIO > MASt3R > DUSt3R. However, they show significantly different rating in **T**exture mode, with MAE > SAM > MASt3R. In the **A**ll mode, MASt3R and DUSt3R achieve the best score, followed by DINO. In Tab. 5, Stable Diffusion (SD) performs the worst in most metrics, Fig. 6-A shows its significant color drift and broken geometry, check *Sup.Mat.* for more qualitative results of geometry. This aligns with the conclusion about SD in Probe3D [22]. Large viewpoint changes cause inconsistency in the feature space (see Fig. 7b). In the following sections, we provide a comprehensive analysis of the insights behind the above ratings.

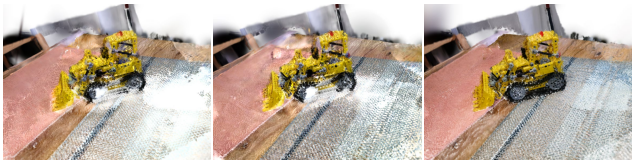


Figure 4. **Texture Blurriness Comparison.** MAE preserves sharper texture over RADIO. IUVRGB is shown for reference.



Figure 5. **MAE *vs.* SD on Texture Awareness.** While both MAE and SD are trained to reconstruct images (MAE in pixel space with an MSE loss and SD in feature space with a denoising loss), SD tends to result in a significant color shift.

Texture-unfriendly Training Strategies. As shown in Tab. 5 and Fig. 6, VFM features perform poorly in **T**exture mode, even worse than the simple IUVRGB encoding shown in Fig. 4. It suggests that current VFM features lack texture awareness, as noted in [82, 106]. One likely explanation for this is that VFMs are often trained for semantic understanding or 3D estimation, which require texture-invariant features to avoid shortcuts [30, 31]. For example, DUSt3R is trained to be texture-invariant for better 3D robustness on diverse in-the-wild captures. Heavy data augmentations in SSL (*i.e.*, DINO [9], BYOL [32], SimCLR [14]), such as color jittering, Gaussian blur, and solarization, encourage the model to produce consistent outputs despite changes in appearance or lightning. Since CLIP is trained on weakly aligned image-text pairs, it often includes ambiguous and coarse semantics that are not discriminative enough to model low-level visual patterns, like colors, materials, and textures [93]. RADIO distills DINO and CLIP, achieving excellent geometry awareness, but also inherits their poor texture awareness (see Fig. 2 and Fig. 4).

	LLFF									DL3DV									Casual								
	Geometry			Texture			All			Geometry			Texture			All			Geometry			Texture			All		
Feature	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DUSt3R	19.88	.7442	.2123	19.01	.7120	.2262	19.87	.7190	.2691	19.64	.7338	.3196	18.01	.6815	.3219	19.39	.7360	.3458	19.29	.6562	.3580	17.54	.5693	.3750	19.19	.6556	.4050
MASt3R	19.89	.7447	.2123	19.01	.7115	.2261	19.99	.7250	.2657	19.64	.7334	.3188	18.07	.6813	.3211	19.41	.7373	.3464	19.30	.6550	.3576	17.59	.5708	.3722	19.37	.6588	.4027
MiDaS	19.81	.7420	.2154	19.00	.7129	.2261	19.86	.7142	.2733	19.47	.7271	.3311	17.94	.6796	.3224	19.22	.7291	.3493	19.24	.6545	.3612	17.52	.5693	.3757	18.96	.6516	.4073
DINOv2	19.77	.7345	.2226	19.04	.7133	.2254	19.91	.7163	.2637	19.47	.7293	.3288	18.00	.6805	.3223	19.27	.7317	.3479	19.42	.6524	.3698	17.64	.5701	.3754	19.21	.6535	.4023
DINO	19.81	.7423	.2140	18.98	.7121	.2260	19.97	.7212	.2744	19.60	.7324	.3209	17.97	.6790	.3219	19.41	.7359	.3476	19.24	.6513	.3614	17.50	.5683	.3756	19.10	.6566	.4056
SAM	19.72	.7354	.2181	18.98	.7133	.2260	19.76	.7144	.2629	19.48	.7297	.3271	17.97	.6822	.3218	19.20	.7272	.3459	19.32	.6469	.3704	17.52	.5725	.3736	19.19	.6569	.3981
CLIP	19.78	.7378	.2221	19.02	.7113	.2276	19.74	.7136	.2822	19.53	.7295	.3304	18.05	.6771	.3235	19.22	.7310	.3563	19.21	.6552	.3719	17.46	.5669	.3743	19.05	.6582	.4084
RADIO	19.73	.7402	.2207	19.06	.7101	.2301	19.56	.6999	.3252	19.48	.7313	.3139	18.03	.6748	.3254	19.20	.7316	.3654	19.54	.6545	.3465	17.52	.5666	.3748	18.67	.6533	.4216
MAE	19.75	.7363	.2183	19.00	.7128	.2249	19.92	.7209	.2612	19.54	.7288	.3248	17.98	.6821	.3207	19.34	.7310	.3448	19.03	.6502	.3690	17.51	.5691	.3758	19.18	.6547	.3974
SD	19.62	.7293	.2234	18.85	.7100	.2297	19.78	.7121	.2656	19.31	.7251	.3276	17.79	.6784	.3260	19.10	.7282	.3500	19.24	.6483	.3649	17.38	.5698	.3789	18.86	.6505	.4053
IUVRGB	15.55	.5765	.3986	19.75	.7303	.2262	15.38	.6175	.4308	14.78	.6326	.4541	18.75	.7023	.3250	14.05	.6431	.4386	13.17	.5454	.5248	17.88	.5927	.3846	13.71	.5917	.4955
	MipNeRF 360									MVImgNet									Tanks and Temples (T&T)								
	Geometry			Texture			All			Geometry			Texture			All			Geometry			Texture			All		
Feature	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DUSt3R	20.82	.5008	.3795	19.10	.4489	.3816	21.02	.5048	.4752	19.47	.6004	.3073	16.88	.5348	.3334	19.43	.5937	.3674	18.85	.6458	.3715	17.53	.6222	.3328	18.61	.6477	.4023
MASt3R	20.92	.5093	.3745	19.21	.4540	.3803	20.92	.5054	.4749	19.49	.6008	.3032	16.91	.5350	.3337	19.49	.5983	.3637	18.80	.6428	.3703	17.68	.6238	.3319	18.76	.6512	.3991
MiDaS	20.89	.5059	.3815	19.05	.4509	.3813	20.84	.5004	.4795	19.35	.5900	.3222	16.82	.5336	.3343	19.34	.5910	.3672	18.53	.6374	.3798	17.64	.6238	.3333	18.32	.6428	.4039
DINOv2	20.81	.4946	.3953	19.05	.4495	.3821	20.75	.4924	.4684	19.35	.5896	.3246	16.88	.5359	.3344	19.43	.5943	.3674	18.71	.6432	.3772	17.58	.6214	.3348	18.43	.6443	.4064
DINO	20.91	.5054	.3769	19.18	.4545	.3795	20.83	.5010	.4772	19.44	.5982	.3071	16.90	.5394	.3329	19.41	.5952	.3683	18.75	.6416	.3733	17.66	.6233	.3330	18.61	.6467	.4030
SAM	20.73	.4913	.3945	19.14	.4556	.3775	20.75	.4949	.4639	19.23	.5899	.3188	16.84	.5346	.3346	19.29	.5915	.3649	18.65	.6421	.3780	17.49	.6217	.3338	18.43	.6425	.4029
CLIP	20.80	.4982	.3913	19.28	.4543	.3807	20.88	.4984	.4773	19.41	.5945	.3098	16.96	.5362	.3358	19.37	.5969	.3695	18.92	.6463	.3729	17.81	.6226	.3316	18.75	.6515	.4052
RADIO	20.87	.5100	.3620	19.35	.4550	.3819	20.91	.5067	.5127	19.54	.6105	.2949	16.99	.5373	.3366	19.60	.5955	.3946	19.19	.6612	.3480	17.84	.6225	.3321	19.01	.6574	.4109
MAE	20.82	.4992	.3884	19.14	.4572	.3781	20.79	.4995	.4668	19.23	.5909	.3142	16.84	.5355	.3328	19.25	.5914	.3680	18.65	.6395	.3758	17.55	.6234	.3333	18.49	.6451	.4000
SD	20.71	.4962	.3985	18.89	.4472	.3839	20.59	.4929	.4672	19.08	.5881	.3185	16.63	.5313	.3389	19.06	.5838	.3660	18.69	.6422	.3772	17.32	.6217	.3374	18.55	.6467	.4020
IUVRGB	16.45	.4075	.5910	19.96	.4797	.3911	16.41	.4187	.5929	14.83	.5069	.4648	17.84	.5568	.3431	15.38	.5362	.4699	15.29	.5846	.4736	18.60	.6526	.3396	15.17	.5948	.4718

Table 5. **Quantitative Results.** We evaluate geometry and texture awareness of VFMs on NVS using Geometry, Texture, and All probing modes. Results indicate that VFM performance varies across datasets, highlighting the importance of dataset diversity. The lack of texture awareness in VFMs limits both **T**exture mode and **A**ll mode, especially in LPIPS. Performance is ranked by color, from worst to best.



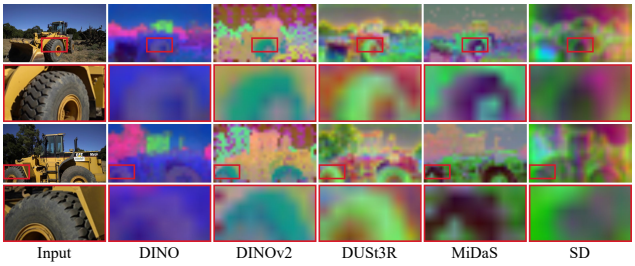
Figure 6. **Qualitative Examples.** We compare novel view renderings across VFM features. In **G**eometry mode, the multi-teacher-distillation method (RADIO) and point-regression-based methods (MASt3R, DUSt3R) produce more plausible geometry, *e.g.*, vehicle front and the wheel, indicating better multi-view consistency. All VFM features struggle in **T**exture mode, and renderings in the **A**ll mode are notably blurred, both reflecting the limited texture awareness of current VFMs.

Texture Benefits from Masked Image Reconstruction. Table 5 shows that the **A**ll mode is impeded by **T**exture mode, leading to worse performance in LPIPS (by an average of $+0.05$) than **G**eometry mode, which does not use VFM features for 3DGS color regression. Terrible texture awareness prevents RADIO from being versatile, as Fig. 2 shows. Visually, as displayed in Fig. 6, novel view renderings in the **A**ll mode tends to appear blurred. Figure R.3 also shows that **T**exture mode, which excludes VFM features for 3DGS geometry regression, exhibits broken structures, while **A**ll mode appears more blurred than the same regions in **G**eometry mode because **A**ll mode relies on VFM features for 3DGS color regression, unlike **G**eometry mode, which freely optimizes colors. To further analyze the mutual correlation of GTA modes, we compute their correlation matrix using average metrics across all datasets, as shown in *Sup.Mat.*'s Fig. R.2. Results indicate that the **A**ll mode is more strongly correlated with **G**eometry mode in PSNR and SSIM, which primarily reflect structural consistency, but is more closely related to **T**exture mode in LPIPS, a metric used to evaluate image sharpness. This further supports the notion that the blurriness observed in the **A**ll mode stems from a lack of texture awareness in VFMs. Texture is obviously crucial for photorealism. How can it be retained in VFMs? As illustrated in Fig. 2, VFMs with masked-image-reconstruction pre-training (*i.e.*, MAE, MASt3R, DUSt3R) rank top in T-LPIPS, and Fig. 4 backs this. MAE's ability to recover sharp textures might be attributed to using only *cropping-only* augmentation. Color jittering degrades results, so it's not included [34]. Additionally, denoise-based image reconstruction leads to color shift, as shown in Fig. 5.

Geometry Benefits from 3D Data. In Fig. 2, RADIO, MASt3R, DUSt3R and DINO rank among the top-4 in geometry awareness metrics. In Fig. 7a, these four features help reconstruct a more complete structure of the excavator, whereas others result in floating artifacts and distortion.



(a) NVS Comparison on Geometry Awareness

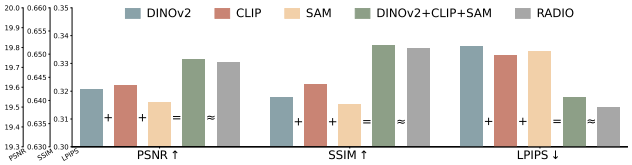


(b) VFM Features from Training Views

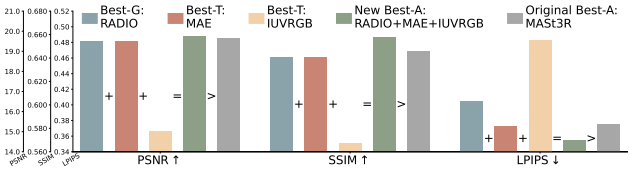
Figure 7. **Feature Consistency Reflected in NVS.** (a) RADIO, MASt3R, DUSt3R, and DINO effectively capture geometry. (b) DINO is consistent across training views, but PE artifacts appear in DINOv2. 3D data proves beneficial, as DUSt3R is consistent; however, MiDaS shows inconsistency, suggesting that pointmap representation is more reasonable than depth. SD also exhibits inconsistency. These inconsistencies lead to poor NVS.

Better geometry awareness implies stronger cross-view consistency, which is also supported by Fig. 7b. *What is the key ingredient to achieve geometry awareness? One crucial factor is 3D data.* Both of MASt3R and DUSt3R are trained with pointmap. What about 2.5D data, like depth or normal maps? It is much worse, see DUSt3R *vs.* MiDaS at Fig. 7a. Please note that, MiDaS and DUSt3R shares the same ViT-L/16 encoder architecture (see Tab. 2) and comparable training scales (3M *vs.* 2M). The depth map estimation may cause inconsistent features for the same object when viewed from different distances. In contrast, pointmap regression [95] encourages the network to generate consistent features across views, as the scene coordinates remain unchanged when the view changes [8, 80].

Model Ensembling Help. RADIO, distilling DINOv2, CLIP and SAM into a single model, achieves the best geometry awareness, as shown in Figs. 2, 6 and 7a. A natural question arises: *Could simply concatenating these features yield comparable results? Yes!* Specifically, we concatenate features of DINOv2, CLIP and SAM, and then apply PCA to reduce feature channels to 256, keeping the size of the network unchanged for a fair comparison. Figure 8a shows that, in **G**eometry mode, feature concatenation (DINOv2+CLIP+SAM) outperforms model distillation (RADIO). This inspires us to further explore: *What*



(a) Concatenated Features *vs.* RADIO



(b) Concatenated Features *vs.* MASt3R

Figure 8. **Feature Concatenation.** (a) RADIO, distilling DINOv2, CLIP, and SAM, achieves superior geometry awareness. **G**eometry mode with concatenated features from the three yields results comparable to RADIO. (b) The **A**ll mode using concatenated VFM features from best **G**eometry mode and **T**exture mode, outperforms the original best **A**ll mode. +: Feature concatenation; ≈: Comparable performance; >: Superior performance.

*if we combine the best **G**eometry mode feature and the best **T**exture mode feature?* *Sup.Mat.*'s Fig. R.2 has indicated that the optimal **A**ll mode should have no weakness in either texture or geometry. As shown in Fig. 8b, the **A**ll mode using concatenated features from RADIO (best in **G**eometry mode) with MAE and IUVRGB (best in **T**exture mode), outperforms the original best **A**ll mode with MASt3R features. This exploration shows the potential of our probing method.

5. Application

Feature Pickup. Inspired by Sec. 4.3, we make three Feat2GS variants to compare with InstantSplat [23], in the NVS task using casual (sparse and uncalibrated) images. Specifically, we pick up the TOP1 of **G**eometry mode, RADIO, as the first baseline. As shown in Tab. 6, Feat2GS with RADIO features achieves better PSNR and SSIM over InstantSplat. The qualitative results in Fig. 9 show that InstantSplat often produces broken structures and discontinuity artifacts. This occurs because optimizing millions of 3DGS for sparse viewpoints leads to overfitting high-frequency details, resulting in suboptimal parameters. In contrast, **G**eometry mode of Feat2GS with RADIO features can produce higher-quality synthesized novel views, thanks to RADIO's strong geometry awareness. This allows us to read out 3DGS from the deep features using a very lightweight (2-layer) MLP, which is crucial for avoiding overfitting.

Feature Ensembling. Since simple feature concatenation is effective (see Fig. 8), it makes sense to consider a straightforward approach: concatenating different VFM features. However, exploring all possible feature combinations is impractical. Thus, we rank the features based on **G**eometry mode performance, and concatenate them in descending (from



Figure 9. **NVS from Casual (Sparse and Uncalibrated) Images.** We compare our **G**eometry mode baselines against InstantSplat, which overfits to training views and results in broken structures. Feat2GS with RADIO produces more consistent results due to alignment from features and compact readout. Concatenating all VFM features leads to more complete geometry, *e.g.*, auditorium seat backs (X-shaped gaps *vs.* straight-line gaps), while fine-tuned DUSt3R* features further refine details, *e.g.*, signage.

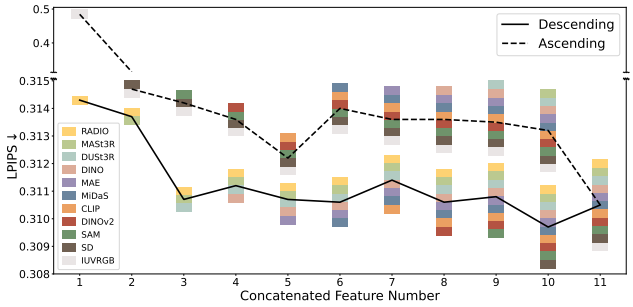


Figure 10. **Ranking-Ordered Feature Concatenation.** Based on performance on **G**eometry mode, VFM features are ranked and concatenated in two orders: descending (best to worst) and ascending (worst to best). Performance improves with more concatenated features and with higher-ranking features.

best- to worst-performing) and ascending (from worst- to best-performing) orders, followed by PCA to reduce the feature dimensions to 256. The results are in Fig. 10. The curve indicates that performance improves as more VFM features are concatenated, with additional gains when higher-ranking VFM features are merged. Based on this observation, we simply concatenate all VFM features as our second baseline. Compared to using only RADIO features, as shown in Fig. 9, "Feat2GS w/ concat all" yields better results. For example, with RADIO features, the auditorium seat backs show X-shaped gaps, whereas concat-all model correctly recover the straight-line gaps between them. Quantitative Tab. 6 also show improvements compared to Feat2GS based on RADIO features, yet it still falls short of InstantSplat in terms of LPIPS. This limitation mainly arises from the low-resolution features extracted from VFM encoders [26, 85]. Though the feature upsampler [26] is leveraged to improve the spatial resolution of features, it does not bring much benefits (detailed in *Sup.Mat.*).

Feature Finetuning. Lastly, we explore whether feature fine-tuning during the warm-start stage is beneficial. The results after fine-tuning different VFM features showed minimal differences, indicating that fine-tuning is effective with any well-initialized features (detailed in *Sup.Mat.*). Since we use DUSt3R to warm-start the readout layer, for simplicity, we compare vanilla DUSt3R with fine-tuned DUSt3R* at Tab. 6. Feature fine-tuning could improve NVS. Figure 9 demonstrates a clear improvement over SOTA InstantSplat.

	All Datasets		
Method	PSNR ↑	SSIM ↑	LPIPS ↓
InstantSplat [23]	18.87	0.6044	0.3039
Feat2GS w/ RADIO	19.73	0.6513	0.3143
Feat2GS w/ concat all	**19.80**	0.6545	0.3105
Feat2GS w/ DUSt3R	19.66	0.6469	0.3247
Feat2GS w/ DUSt3R*	19.75	**0.6561**	**0.2928**

Table 6. **Baselines of Feat2GS** in NVS from casual (sparse and uncalibrated) images. We compare **G**eometry mode with RADIO features, concatenation of all VFM features, DUSt3R, and fine-tuned DUSt3R* features against the current SOTA InstantSplat.

6. Conclusion

We now return to our original question: Are visual foundation models (VFMs) aware of geometry and texture? To give an answer based on diverse datasets, we proposed Feat2GS, a method that maps features of VFMs to 3DGS, allowing us to explore their geometry and texture awareness through 2D images without requiring 3D ground-truth.

Our work reveals new insights: correlations between novel view synthesis and 3D metrics (Tab. 4, Fig. 3), color shift in SD (Fig. 5), and better view consistency from models trained on pointmaps over depth (Fig. 7). In addition, Feat2GS effectively harnesses VFMs for novel view synthesis task on sparse casual captures. These findings suggest that predicting 3D Gaussians from various views in a canonical space and training the model with photometric loss is a promising strategy for developing 3D VFMs, also noted at [24, 106]. Moreover, VFM feature ensembling is an interesting topic worth exploring [73, 79], and we demonstrate that this can be effectively achieved in Feat2GS through simple concatenation. We hope these insights, along with Feat2GS— a versatile tool for future model exploration — will advance VFM research and drive progress in 3D vision.

Acknowledgments. We thank *Yuxuan Xue*, *Vladimir Guzov*, *Garvita Tiwari* for their valuable feedback, and the members of *Endless AI Lab* and *Real Virtual Humans* for their help and discussions. This work is funded by the Research Center for Industries of the Future (RCIF) at Westlake University, the Westlake Education Foundation, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (EmmyNoether Programme, project: Real Virtual Humans), and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. *Yuliang Xiu* also received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.860768 (CLIPE project), and Max Planck Institute for Intelligent Systems. *Gerard Pons-Moll* is a Professor at the University of Tübingen endowed by the Carl Zeiss Foundation, at the Department of Computer Science and a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016. 4

[2] Görkay Aydemir, Weidi Xie, and Fatma Güney. Can visual foundation models achieve long-term point tracking? *arXiv preprint arXiv:2408.13575*, 2024. 2, 3

[3] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 2, 4

[5] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4160–4169, 2023. 3, 4

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1, 2

[7] Tyler Bonnen, Stephanie Fu, Yutong Bai, Thomas O'Connell, Yoni Friedman, Nancy Kanwisher, Joshua B. Tenenbaum, and Alexei A. Efros. Evaluating multiview object consistency in humans and image models, 2024. 3

[8] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *ECCV*, 2024. 2, 7, 17

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 2, 4, 5, 6, 15, 17

[10] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 1, 3

[11] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[12] Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[13] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 1, 2

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 5

[15] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12952, 2022. 17

[16] Yue Chen, Xingyu Chen, Xuan Wang, Qi Zhang, Yu Guo, Ying Shan, and Fei Wang. Local-to-global registration for bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8264–8273, 2023. 3

[17] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *ECCV*, 2024. 3

[18] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. 17

[19] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 16, 19

[20] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 3

[21] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1033–1038. IEEE, 1999. 2

[22] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 2, 3, 5, 16, 17

[23] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2024. 2, 3, 4, 7, 8

[24] Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, et al. Large spatial model: End-to-end unposed images to semantic 3d. *NeurIPS*, 2024. 2, 3, 8

[25] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2093–2103, 2024. 2

[26] Stephanie Fu, Mark Hamilton, Laura E. Brandt, Axel Feldmann, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution. In *The Twelfth International Conference on Learning Representations*, 2024. 8, 15, 16, 19

[27] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. 2024. 3

[28] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, 2024. 2

[29] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20796–20805, 2024. 3, 4

[30] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 5

[31] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 5

[32] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 5

[33] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011. 3

[34] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 4, 5, 6, 15, 17

[35] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *ICLR*, 2024. 1, 2, 3

[36] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 2

[37] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5885–5894, 2021. 3

[38] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5846–5854, 2021. 3

[39] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias, 2024. 3

[40] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 1, 2

[41] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 3, 15

[42] Justin* Kerr, Chung Min* Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023. 15

[43] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *ECCV*, 2024. 2

[44] Chung Min Kim, Mingxuan Wu, Justin Kerr, Ken Goldberg, Matthew Tancik, and Angjoo Kanazawa. Garfield: Group anything with radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21530–21539, 2024. 15

[45] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2, 4, 5, 6, 15, 17

[46] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 16, 19

[47] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):1–13, 2017. 2, 4

[48] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Advances in Neural Information Processing Systems*, 2022. 15

[49] Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *arXiv preprint arXiv:2407.08447*, 2024. 17

[50] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 2, 3, 4, 5, 6, 15, 17

[51] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 15

[52] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5741–5751, 2021. 3, 4

[53] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 2, 4

[54] Drew Linsley, Peisen Zhou, Alekh Karkada Ashok, Akash Nagaraj, Gaurav Gaonkar, Francis E Lewis, Zygmunt Pizlo, and Thomas Serre. The 3d-pc: a benchmark for visual perspective taking in humans and machines. *arXiv preprint arXiv:2406.04138*, 2024. 3

[55] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 3

[56] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 17

[57] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36:655–677, 2023. 3

[58] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021. 17

[59] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. GNeRF: GAN-based Neural Radiance Field without Posed Camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[60] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16539–16548, 2023. 3

[61] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 2, 4

[62] 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, 15

[63] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 3

[64] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 3

[65] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 5, 6, 15, 17

[66] Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomáš Hodaň. Foundpose: Unseen object pose estimation with foundation features. *ECCV*, 2024. 2, 17

[67] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 15

[68] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023. 15

[69] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 15

[70] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 4, 5, 6, 15, 17

[71] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 1, 2, 4, 5, 6, 15, 17

[72] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 16

[73] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12490–12500, 2024. 1, 4, 5, 6, 8, 15, 17

[74] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 15

[75] Arijit Ray, Dina Bashkirova, Reuben Tan, Kuo-Hao Zeng, Bryan A Plummer, Ranjay Krishna, and Kate Saenko. R2d3: Imparting spatial reasoning by reconstructing 3d scenes from 2d images. 2, 3

[76] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 4, 5, 6, 15, 17

[77] 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 image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9420–9429, 2024. 3

[78] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[79] Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. In *8th Annual Conference on Robot Learning*, 2024. 3, 8

[80] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 7

[81] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023. 15

[82] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024. 2, 3, 4, 5

[83] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. *arXiv preprint arXiv:2404.15259*, 2024. 2

[84] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 17

[85] Saksham Suri, Matthew Walmer, Kamal Gupta, and Abhinav Shrivastava. Lift: A surprisingly simple lightweight feature transform for dense vit descriptors. In *European Conference on Computer Vision*, pages 110–128. Springer, 2025. 8

[86] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 3

[87] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4190–4200, 2023. 3

[88] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural Feature Fusion Fields: 3D distillation of self-supervised 2D image representations. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2022. 15

[89] Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[90] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *ECCV*, 2024. 3, 17

[91] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. *arXiv preprint arXiv:2112.05139*, 2021. 15

[92] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 3, 4, 17

[93] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving gener-

ation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 5

[94] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv e-prints*, pages arXiv–2407, 2024. 17

[95] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 1, 2, 3, 4, 5, 6, 7, 15, 17

[96] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 3, 4

[97] Frederik Warburg*, Ethan Weber*, Matthew Tancik, Aleksander Hołyński, and Angjoo Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[98] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 2, 17

[99] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21551–21561, 2024. 3

[100] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, pages 641–676. Wiley Online Library, 2022. 15

[101] Congrong Xu, Justin Kerr, and Angjoo Kanazawa. Splatfacto-w: A nerfstudio implementation of gaussian splatting for unconstrained photo collections. *arXiv preprint arXiv:2407.12306*, 2024. 17

[102] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. 2022. 3

[103] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8254–8263, 2023. 3

[104] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. In *International Conference on Learning Representations*, 2024. 16

[105] Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas J. Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Dvt: Denoising vision transformers. 2024. 16, 17

[106] 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, 2024. 2, 3, 5, 8

[107] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 2024. 1, 2

[108] Jianglong Ye, Naiyan Wang, and Xiaolong Wang. Featurenerf: Learning generalizable nerfs by distilling foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8962–8973, 2023. 15

[109] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision*, pages 162–179. Springer, 2025. 15

[110] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for Gaussian splatting. *arXiv preprint arXiv:2409.06765*, 2024. 15

[111] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 3

[112] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023. 2, 4

[113] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2D Feature Representations by 3D-Aware Fine-Tuning. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 17

[114] Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In *European Conference on Computer Vision*, pages 341–359. Springer, 2025. 17

[115] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024. 3, 17

[116] Xiaoshuai Zhang, Zhicheng Wang, Howard Zhou, Soham Ghosh, Danushen Gnanapragasam, Varun Jampani, Hao Su, and Leonidas Guibas. Condense: Consistent 2d/3d pre-training for dense and sparse features from multi-view images. In *European Conference on Computer Vision*. Springer, 2024. 17

[117] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 15

[118] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 15

[119] Haoyi Zhu, Honghui Yang, Yating Wang, Jiange Yang, Limin Wang, and Tong He. Spa: 3d spatial-awareness enables effective embodied representation. *arXiv preprint arXiv:2410.08208*, 2024. 3

[120] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting, 2023. 3

[121] Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. *arXiv preprint 2410.12781*, 2024. 3

[122] Yiming Zuo, Karhan Kayan, Maggie Wang, Kevin Jeon, Jia Deng, and Thomas L Griffiths. Towards foundation models for 3d vision: How close are we? *arXiv preprint arXiv:2410.10799*, 2024. 3