

Zero-Shot Novel Depth Synthesis Using 3D Foundation Models Scene Representations

Denis M. Akola¹ and David F. Fouhey^{1,2}

¹ New York University, Tandon School of Engineering, 1 MetroTech Center,
Brooklyn, NY 11201, USA

² Courant Institute of Mathematical Sciences, 251 Mercer St, New York, NY 10012

Abstract. 3D Foundation Models (3DFMs) such as VGGT have recently pushed the boundaries of 3D vision by predicting rich unified representations with feed-forward transformers. The scene representations learned by these models enable strong performance on multiple 3D vision tasks. In this paper, we investigate using their internal representations to infer 3D in the scene from new views. Our hypothesis is that in order to solve the task of 3D reconstruction, these models need to learn a representation that includes a large amount of general knowledge about 3D scenes. After showing that it is possible to decode hidden surfaces from internal 3DFM representations, we propose a method, Z3D, that estimates pointmaps in unseen views by doing latent diffusion on 3DFM representation. We show that Z3D can predict realistic depthmaps for new views across multiple datasets.

1 Introduction

Understanding the full 3D structure of a scene from limited observations remains a fundamental challenge in computer vision. In many real-world settings like robot navigation, AR/VR and autonomous systems, we only observe partial views of a scene, yet must reason about the geometry that lies outside the visible surface, including regions occluded by foreground objects. Novel depth synthesis addresses this challenge. Given one or a few input views, the goal is to predict plausible and geometrically consistent depth for unseen viewpoints, including content hidden behind objects. Beyond view rendering, this capability enables physical interaction, planning, and spatial reasoning in incomplete environments, where missing geometry must be inferred rather than directly observed.

Recent advances in neural scene representations such as NeRF [24], 3D Gaussian Splatting [20] and other neural methods [10, 12, 13, 18] have significantly improved novel view synthesis by optimizing continuous 3D representations from multi-view imagery. However, these methods primarily interpolate between observed views and are not designed to hallucinate geometry behind occlusions, especially in sparse-view settings. Prior works [10, 11] on novel depth synthesis often jointly learn view and depth prediction from multi-view data, but progress toward generalizable depth synthesis from sparse inputs remains limited. More recently, generative models [12, 19] have been explored for depth prediction, but

these often operate in 2D image space and lack strong 3D consistency priors, limiting their ability to produce coherent geometry across viewpoints. The core challenges stem from geometric ambiguity under limited observations and the difficulty of learning representations that transfer across diverse scenes.

The key insight of this paper is that the internal representations of pretrained 3D foundation models (3DFMs) [23, 45, 49, 52] can be effectively combined with diffusion-based models for generalizable novel-view depth synthesis. Rather than treating novel depth synthesis as pure reconstruction or view interpolation, we formulate it as a conditional generation task over scene geometry. Specifically, we use the intermediate representations of 3DFMs for conditioning as well as the latent space in which to do latent diffusion [32]. After diffusion, the predicted latents can be converted to geometry by the original decoder. By decoupling representation learning (which is taken care of by the 3DFM) from generative completion (handled by diffusion), our approach enables plausible hallucination of occluded or unobserved structure, and encourages multi-view consistency. To further support our premise that 3DFMs encode geometry beyond directly visible surfaces, we show that simply attaching a lightweight decoder to their features enables inference of hidden surfaces, demonstrating that the learned scene representations are well-suited for reasoning about occluded geometry.

We propose **Z3D—Zero-Shot 3D Depth**, a simple and effective method that leverages the rich geometric scene representations learned by 3DFMs in combination with a diffusion framework to generate novel-view depth maps. Z3D integrates these pretrained 3DFM representations with the relative pose of a novel view to predict consistent and accurate depth for novel view cameras. We evaluate our method on multiple datasets. We compare Z3D with state-of-art methods [18] for novel view depth synthesis.

2 Related Work

We aim to generate consistent depth maps for novel views of a scene, given a sparse set of images (2 or more views). We first infer a geometric scene representation from 3DFMs and, using a diffusion model conditioned on this representation and the poses of the target views, predict novel-view depths.

Novel View and Depth Synthesis. NeRF [24] and its extensions [25, 26, 54] render novel views of observed surfaces but often require many input images and per-scene optimization. In sparse view settings, they tend to produce incomplete geometry and cannot infer surfaces outside the original field of view or behind occlusions. In contrast, our method infers novel-view geometry from sparse inputs without per-scene optimization. Methods such as PixelNeRF [54], SRT [36], RUST [35], LVSM [18], and RayZer [17] generalize across scenes but focus on photometric quality rather than accurate depth, limiting their ability to hallucinate unseen structure. Our approach explicitly models geometry, generating consistent depth maps beyond visible surfaces. CUT3R [48] is a stateful 3D reconstruction model capable of recovering scene geometry from RGB observations, predicting depth for novel views, and inferring unobserved scene regions

by querying virtual camera viewpoints. Unlike CUT3R, our method incorporates pre-trained 3DFM representations within a diffusion framework. This design leverages the structured geometric knowledge learned by 3DFM while benefiting from the expressive generative power of diffusion models, resulting in more accurate and realistic novel-view depth synthesis.

A few works jointly predict novel views and depth [10, 11]. For example, [10] learns scene representations within the diffusion model and predicts depth for a single target view under restrictive camera assumptions. In contrast, we use pretrained 3DFM features, decoupling representation from generation, and produce plausible depths for multiple novel viewpoints without assuming that the target camera(s) are always placed at the origin.

3D Foundation Models (3DFMs). Traditional 3D reconstruction methods such as Structure-from-Motion (SfM) [2, 14, 40] and Multi-View Stereo (MVS) [7–9] recover camera poses and geometry through local feature matching and joint optimization via bundle adjustment. Recent 3D foundation models instead predict camera poses, depth, and point maps in a single feed-forward pass from multiple input images. Representative works in this line of research include DUST3R [49] and its extensions [16, 21, 48, 52], as well as multiview models such as VGGT [45] and WorldMirror [23]. These models are not designed for novel-view synthesis out of the box, but we repurpose their representations for this task by pairing them with diffusion models. Specifically, we treat their internal features as strong scene-level geometry representation. We empirically show that a lightweight decoder can use these features to predict hidden surfaces (see § 3.1), demonstrating that 3DFMs encode information useful for reasoning about occluded geometry and unseen viewpoints.

Diffusion Models. Denoising Diffusion Probabilistic Models (DDPMs) [15] generate samples by iteratively reversing a diffusion process, which involves denoising Gaussian noise into structured outputs. Conditional DDPMs incorporate information like text [34], reference images [33], or semantic maps [57] to guide generation. Latent Diffusion Models (LDMs) [6, 28, 32] perform denoising in a compact latent space. Diffusion has recently been applied to monocular depth [12, 19, 38, 39, 51, 59] and camera pose [47, 56] prediction. These models operate in a compact latent space similar to original LDMs, allowing standard noise schedulers to work out of the box. These approaches [12, 19, 38, 39, 51, 59] predict depth only for surfaces explicitly visible in the input images and cannot hallucinate occluded or unobserved geometry. In contrast, our method conditions diffusion on high dimensional geometric features extracted from pretrained 3DFMs that requires careful tuning of noise schedulers.

3 Method

Our goal is to predict the depth of a novel view given one or more input images as well as the pose of set of target camera views. In the base case of two views where one view is the source and the other view is target, the model should be able to plausibly hallucinate about the depth of the target view. With

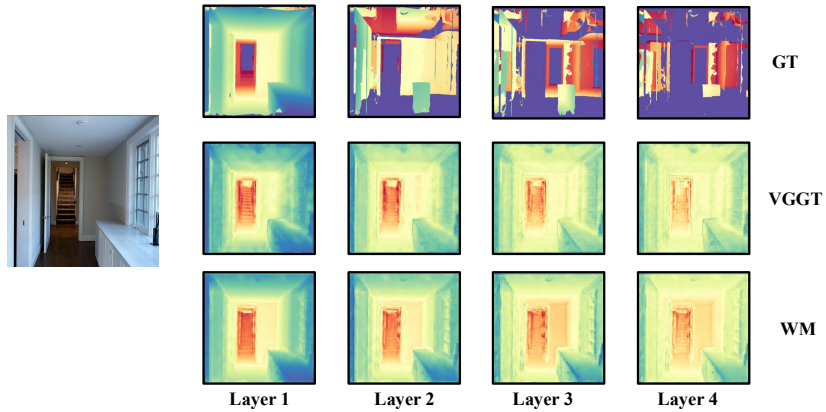


Fig. 1: Linear Probe Layered Depth Image (LDI) prediction capability of 3DFMs. To test whether 3DFMs implicitly contain information about unseen views, we train a linear model to predict LDIs. Good performance on LDI prediction after the first layer suggests that the model understands the full 3D scene, including hidden surfaces. For each example: (left) input RGB; (top) ground-truth LDI layers; (bottom) linearly decoded LDI layers. The ability to recover the non-trivial structure of the scene with only a linear layer suggests that 3DFMs implicitly encode this information.

more views, the model should be able to predict the depth of the target view(s) more accurately. We propose Z3D, a simple and effective framework for novel view depth synthesis that integrates diffusion-based generative modeling with 3DFMs to leverage their complementary strengths. In what follows, we analyze how the scene representations learned by 3DFMs encode geometric structure beyond what is immediately apparent, using a lightweight probe to predict layered depth for each scene. Building upon our findings, we introduce Z3D method that leverages the encoding of a scene from 3DFMs within a diffusion framework to learn to generate consistent depth of novel views.

Background of 3DFMs. 3DFMs [23, 45] encode input images into patch tokens using a shared transformer backbone. Each image is divided into patches, producing tokens that are processed independently in frame attention and jointly across images in global attention. The pretrained patch tokens from the transformer backbone capture rich multi-view geometric information and are fed into task-specific heads, such as depth, point map, or camera pose. In our method, we extract these pretrained patch tokens and use them as a scene representation to condition a diffusion model which allows us to predict the depth for novel viewpoints. These viewpoints reveal occluded and unobserved surfaces.

3.1 Do 3DFMs Know about Unseen Structure?

We first test whether knowledge of this information is already embedded in the 3DFM representations. To accomplish this, we follow the work of Banani et

al. [4] and probe the knowledge of these networks. As our target representation, we train a linear probe to predict layered depth images (LDI) [41], a simple representation of occluded parts of the scene. LDIs generalize depth maps and represent, at each pixel, the first k surfaces that the ray through the pixel passes through. Accurately estimating the 2nd surfaces and subsequent surfaces of an LDI requires knowledge of the occluded parts of the scene.

Table 1: LDI results.

Method	AbsRel $\downarrow$ $\delta < 1.25$ $\uparrow$	
Avg LDI	0.319	0.564
VGGT-LDI	0.197	0.717
WM-LDI	0.167	0.789

We test this by evaluating how well the last layer of a 3DFM’s DPT head can predict LDIs. Specifically, we expand the DPT head of VGGT [45] to produce four depth layers and confidence maps and train the model while keeping all other layers fixed. Figure 1 shows that 3DFM produces plausible layered depth predictions that extend beyond the first visible surface. These results suggest, that 3DFMs do contain implicit knowledge for reasoning about occluded and hidden scene structure. We also quantitatively confirm that linear probes on 3DFM features predict LDIs. We report results with VGGT and WM in Table 1. Despite being a weak linear model, they halve the error rate compared to using the dataset average, showing the hidden/unseen information is embedded in 3DFMs.

3.2 Network Architecture

Our goal is to predict novel-view depth maps by performing conditional diffusion on 3DFM patch tokens, conditioned on relative poses and source view images. Z3D achieves this by combining a pretrained 3DFM with a Pose Encoder, a conditional Diffusion Model, and a DPT head. The pretrained 3DFM provides a high-dimensional latent representation of the scene, capturing rich multi-view geometry. The Pose Encoder encodes the relative camera poses of target views, while the diffusion model predicts the target view scene representation from noise conditioned on both the source views and target poses. Finally, the DPT head decodes the sampled scene representation into depth maps. Conceptually, this setup can be seen as latent diffusion operating in the space of 3DFM patch tokens, enabling hallucination of occluded surfaces and multi-view consistency. During training, both source and target images are available. We first compute scene representations for the source and target views using the frozen 3DFM backbone. Gaussian noise is added to the target view representation, and the diffusion model learns to predict this noise conditioned on the source view representations and relative target view poses. Conditioning is implemented via a cross-attention block in each DiTBlock, while target pose embeddings are added directly to the timestep embeddings, allowing the model to reason about depth from relative camera positions.

3DFM Backbone, ϕ . The 3DFM backbone tokenizes each source image into patches and produces patch tokens, which are processed through alternating frame-level and global attention blocks to capture both per-frame and multi-frame geometry. The resulting patch tokens are concatenated to form a scene

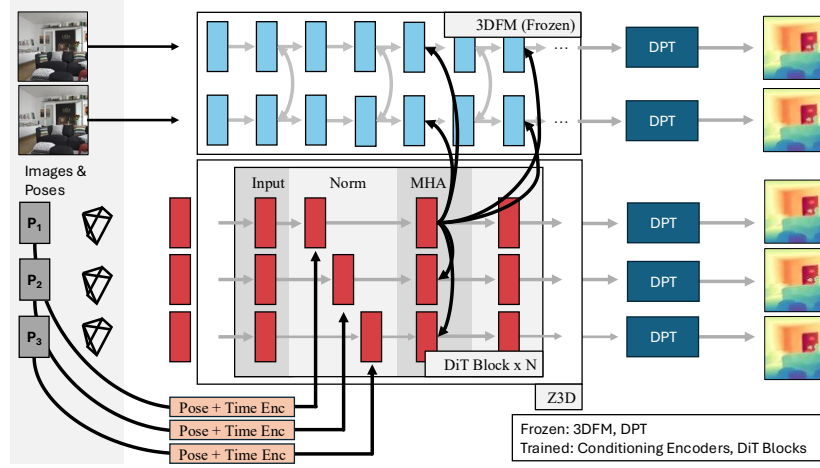


Fig. 2: Z3D Inference Pipeline. Z3D takes one or more input images (top left images in this figure) and multiple query poses (p_1, p_2, p_3 etc). The input images are passed into a 3DFM. Z3D infers the depthmaps at the query poses via latent diffusion using 3DFM features. In each DiT block, Z3D: (i) conditions on the encoded pose (plus diffusion timestep) via adaptive layer normalization following DiT practice; and (ii) attends to the 3DFM intermediate activations while doing multiheaded attention. The diffusion-predicted 3DFM feature is passed directly into the frozen DPT head.

representation for the input images. Only tokens from layer 17 are used for conditioning the diffusion model, leveraging the most informative geometric features while keeping computation efficient.

Pose Encoder, p_θ . Relative camera poses of target views are encoded as 12-dimensional vectors per view. Each pose is Fourier-encoded and processed through an MLP to produce embeddings that match the feature dimension of the 3DFM patch tokens. All poses are represented relative to the first view in the source set, simplifying spatial reasoning for the diffusion model.

Diffusion Model, f_θ . Our diffusion model follows the DiT architecture [28] and operates directly in the high-dimensional latent space of 3DFM patch tokens. Let I_s denote the source view image(s), I_t the target view image(s), and P_t the relative camera pose(s) of the target view(s). We use a shared image backbone ϕ to extract features from both target and source images: $f_t = \phi(I_t)$ and $f_s = \phi(I_s)$. We denote the target representation as $x_0 = f_t$ and the source representations as $z_s = f_s$. Also, we encode the relative pose using an MLP $p = p_\theta(P_t)$.

Recent 3DFM architectures such as VGGT [45] and WorldMirror [23] aggregate transformer tokens from multiple intermediate layers for depth prediction. Each of these token representations consists of high-dimensional embeddings (e.g., 2048-dimensional in both VGGT and WorldMirror), making it computationally prohibitive to apply diffusion directly to the concatenated multi-layer outputs used for depth estimation.

Following an analysis similar to [58], we observe that, for depth prediction, tokens from layer 17 contribute significantly more than those from the other three aggregation layers. Based on this observation, we restrict diffusion to the scene tokens extracted from layer 17 for the target view(s), substantially reducing computational cost while preserving the most depth-relevant representation. To mitigate potential information loss due to this constraint, we use the outputs from all four aggregation layers of the source views as conditioning signals. This allows the model to access richer multi-scale geometric information during cross-attention, while keeping the diffusion process tractable. We provide additional details of this analysis in the supplementary material.

During training, only the target-view tokens x_0 are noised, while the source-view tokens z_s remain clean and serve solely as conditioning signals. The forward diffusion process corrupts the target tokens over timesteps $t \in \{1, \dots, T\}$: $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and $\alpha_t = \prod_{s=1}^t (1 - \beta_s)$, with $\{\beta_t\}_{t=1}^T$ defining the variance schedule.

The denoiser f_θ is trained to predict the velocity of injected noise: $\hat{v}_t = f_\theta(x_t, t, z_s, p)$, where x_t are the noised target tokens. Conditioning on z_s is implemented through cross-attention layers inserted in each DiTBlock, allowing the model to extract geometry most relevant to the target view. The pose embedding p is added to the timestep embedding.

Because standard noise schedulers are typically designed for lower-dimensional latent spaces (e.g., dimension $\lesssim 1024$), directly applying them to the higher-dimensional 3DFM token space (typically ≥ 1024) can lead to instability. Following [60], we apply a timestep shift to the noise scheduler, inspired by [6], which stabilizes training and enables effective denoising in this richer latent space.

Figure 2 shows Z3D’s inference pipeline. We initialize $x_T \sim \mathcal{N}(0, I)$ and iteratively apply f_θ to recover an estimate of the clean target tokens x_0 . The reconstructed tokens are then passed through the pretrained decoder D to produce the final target-view depth map: $\hat{d} = D(x_0)$.

DPT Head, D . The DPT head acts as a decoder, analogous to the decoder in a variational autoencoder (VAE) [27]. It is the depth prediction heads of the 3DFMs. The sampled target-view scene representation tokens, x_0 produced by the diffusion model are passed through the DPT head, which decodes them into depth maps for the target views.

3.3 Implementation Details

We implement Z3D using PyTorch and the modified DiT model outlined in § 3.2). Our diffusion model is trained using a flow-matching noise scheduler with a flow-velocity prediction objective, following the formulation introduced in Stable Diffusion v3 [6]. During training, we apply `FlowMatchEulerDiscreteScheduler` [6] noise scheduler with 1000 timesteps. We adjust the timestep shift factor using a dimension-dependent scaling rule, $\alpha = \sqrt{m/n}$, where the scaling compensates for changes in the latent dimensionality. The timestep shift factor in an Euler flow matching scheduler is a small offset applied to time steps to avoid evaluating

the model exactly at the boundary times (0 and 1), improving numerical stability during sampling. Following [6, 60], we use $n = 4096$ as the reference base dimension and set m to the effective data dimension of the 3DFM backbone representation. At inference time, we apply the `FlowMatchEulerDiscreteScheduler` [60] scheduler and only sample 50 steps. We train our model in two stages. In the first stage, the diffusion model is trained under a one-source-one-target view setting using an effective batch size of 128 for 98k training steps. In the second stage, the model is initialized with the weights obtained from Stage 1 and further trained using a multi-view configuration consisting of two source views and four target views, with an effective batch size of 32 for 156k steps.

Our model is not sensitive to batch size; however, we adopt this incremental training strategy because training the Stage 2 configuration from scratch was observed to be unstable and led to slower convergence. Initializing from Stage 1 led to much stable optimization and faster convergence. We train the model using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, no weight decay, and a learning rate of 2×10^{-4} for both training stages. We apply a linear warmup over the first 10% of the total training steps, followed by a learning rate decay schedule that begins at 30% of the training progress and gradually decreases the learning rate to zero by the end of training.

3.4 Training Datasets

Z3D is trained on sequences of overlapping camera views, typically containing 2–6 images per sequence. The supplement has more data preparation details, but briefly: Training samples are constructed by mining 3D datasets to identify groups of views with sufficient geometric overlap (defined via pairwise camera frustum overlap). To ensure robustness and broad applicability, Z3D is trained on a large-scale aggregation of five diverse datasets that are both real and synthetic. This combined corpus provides extensive coverage across indoor and outdoor environments. The specific datasets include MegaDepth [22], Hypersim [31], Taskonomy [55], Replica [43], and Habitat HM3D [37].

4 Experiments

In this section, we present the experimental framework for Z3D. Our system is designed to predict novel-view depths from sparse input views using 3DFMs. Given the novelty of this problem, there are no existing methods that directly address this setting. To evaluate Z3D, we construct datasets as described in § 3.4 and perform a comprehensive experimental analysis. We consider two recent 3DFMs: VGGT [45] and WorldMirror [23].

4.1 Baselines

Several works attempt to predict novel-view depths jointly with novel-view synthesis [10, 11]. However, to the best of our knowledge, there are no existing

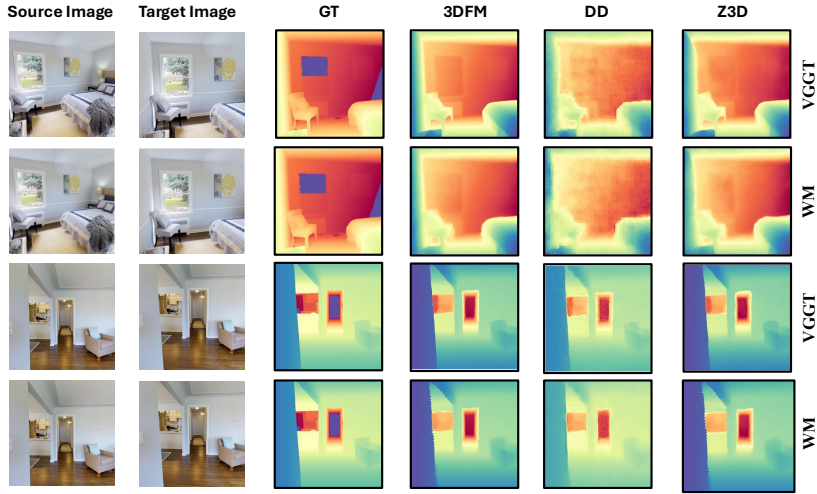


Fig. 3: Qualitative comparison of Novel View Depth Predictions. Columns show: source RGB, target RGB, ground truth depth, 3DFM predicted depth, Depth Diffusion prediction, and Z3D prediction. Z3D results are substantially more smooth as compared to results from a method that works directly in patch-space.

methods that repurpose 3DFMs specifically for this task. To address this, we construct strong baselines from methods that tackle parts of the problem. Our work is closely related to MVGD [10]. Unfortunately, since the code for MVGD [10] is not available, we were unable to perform a direct comparison.

LVSM + 3DFM. We use LVSM [18] as a baseline because it is a state-of-the-art, scene-agnostic novel view synthesis model that generates novel RGB views from sparse observations and camera parameters (intrinsics and extrinsics). Since LVSM does not directly predict depth, we apply pre-trained 3DFMs (e.g., VGGT and WorldMirror) to the synthesized images to estimate depth for the posed cameras. This baseline evaluates the extent to which a general-purpose image synthesis model, combined with a strong monocular 3D foundation model, can be used for novel-view depth prediction. We denote these variants as LVSM+VGGT and LVSM+WM, corresponding to LVSM followed by VGGT and WorldMirror depth estimation, respectively.

Depth Diffusion (DD). As an additional baseline, we modify our approach by training the diffusion model directly in pixel space on the predicted depth maps from 3DFMs, rather than modeling the distribution of the target scene’s latent representation. To separate out the effects of the architecture from the space in which inference is done, we use the Z3D architecture but change the output space. We consider two variants of this baseline: Depth Diffusion (VGGT), hereafter referred to as VGGT-DD, and Depth Diffusion (WorldMirror), hereafter referred to as WM-DD.

4.2 Evaluation

In this section, we describe our quantitative and qualitative evaluations.

Datasets. We follow a similar procedure as outlined in § 3.4 to preprocess datasets for evaluation. We evaluate our model on both indoor and outdoor scenes, covering a diverse set of environments. Specifically, we perform quantitative evaluation on DTU [1], NRGBD [3], 7-Scenes [42], and an in-domain evaluation set sampled from the test splits of training datasets.

Depth Estimation Metrics. Following the affine-invariant depth evaluation protocol [29], we first align the predicted depth maps $\hat{d}$ to the ground truth d using least squares fitting (Weiszfeld method [5]), producing the absolute aligned depth map $a = \hat{d} \times s + t$ in the same units as the ground truth. The depth quality is then assessed using two widely adopted metrics [29, 30, 53]. The first, Absolute Mean Relative Error (AbsRel), is computed as $\text{AbsRel} = \frac{1}{M} \sum_{i=1}^M |a_i - d_i|/d_i$, where M is the total number of pixels. The second metric, $\delta < 1.25$, measures the fraction of pixels satisfying $\max(a_i/d_i, d_i/a_i) < 1.25$.

Multiview 3D Reconstruction Metrics. Since we predict novel depths for multiple cameras in a scene, we evaluate multiview 3D reconstruction by projecting the predicted depthmaps to point clouds. Consistent with prior works [3, 44, 48, 49], we report Accuracy (Acc.), and Completion (Comp.), which together quantify both the fidelity and completeness of the reconstructed 3D geometry.

Qualitative Results. Figure 3 presents qualitative results of Z3D, and Depth Diffusion (DD) baselines. Both Z3D-VGGT and Z3D-WM predict reasonable novel-view depth when conditioned on the target camera pose. However, DD methods yield noticeably less smooth depth maps, likely because they operate directly in depth space rather than a latent representation.

To further substantiate this observation, we visualize the depth gradients of predictions from Z3D and DD baselines in Figure 4. In Figure 4, we observed pronounced gradient spikes in VGGT-DD depths in comparison to Z3D-VGGT, a trend that is consistently observed when comparing WM-DD and Z3D-WM.

We also visualize point clouds obtained by projecting predicted depth maps for novel viewpoints in Figures 5 and 6. In Figure 5, DD-based methods produce noticeably noisier point clouds compared to Z3D predictions. In the multi-view setting (Figure 6), Z3D generates consistent depth across novel views given only a few source views (two in this case).

Quantitative Evaluation We evaluate our method on out-of-domain datasets under two settings: 1-source $\rightarrow$ 1-target and 2-sources $\rightarrow$ 4-targets, to assess its generalization ability across unseen domains. Our approach is flexible with respect to the number of input views and performs well in both few-view and sparse-view settings. Under 2-view setting, our models often outperform our baselines on depth map metrics and are comparable in point cloud reconstruction metrics. However, we qualitatively find that Z3D produces smoother results than the depth diffusion methods (see Figure 4).

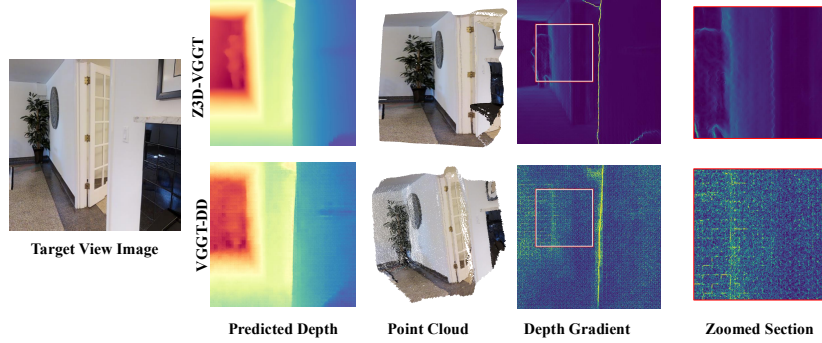


Fig. 4: Depth gradient analysis of Z3D-VGGT and VGGT-DD model predictions

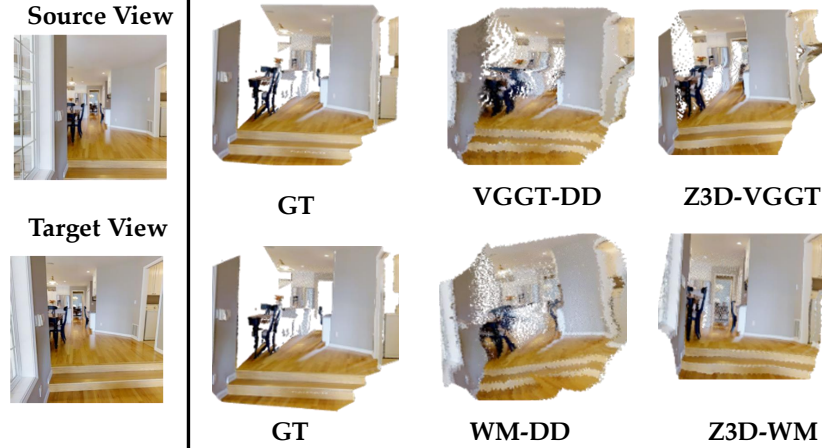


Fig. 5: Qualitative comparison of novel-view point cloud reconstructions from a single source image and target pose. Although VGGT-DD and WM-DD achieves competitive quantitative performance, its reconstructed point clouds exhibit noticeably higher noise and reduced geometric sharpness compared to Z3D-VGGT and Z3D-WM, which produce cleaner and more coherent 3D structures. Top left: source view (left of the vertical line). Bottom left: target view (not observed by the model). First row: Predictions of VGGT , VGGT-DD and Z3D-VGGT respectively. Second row: Predictions of WM , WM-DD and Z3D-WM respectively.

In-domain Evaluation. In this setting, we evaluate Z3D and DD models on subsets of the test splits from the training datasets. As shown in Tables 2, 3, 4, and 5, all models achieve strong performance on their respective training distributions.

Out-of-domain Evaluation. Here, we evaluate our method and baselines on the test splits of datasets not seen during the training of our model and baselines. Our LVSM+3DFM baselines provide a strong comparison since LVSM is trained

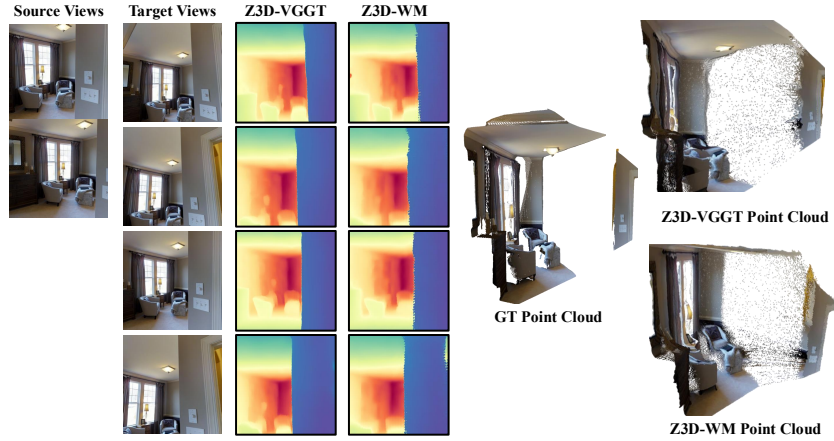


Fig. 6: Novel-view depth prediction from 2 source views $\rightarrow$ 4 target views. Conditioned on the target camera poses, Z3D-VGGT and Z3D-WM generate geometrically consistent depth maps across all target views, resulting in a coherent reconstructed point cloud.

end-to-end for novel-view synthesis and is expected to encode scene geometry accurately. However, we observe that LVSM sometimes struggles to generate novel views that are fully consistent with the target pose, suggesting that optimizing for perceptual quality can come at the expense of faithfully preserving the underlying scene geometry.

From Tables 2 and 3, our Depth Diffusion baselines (WM-DD and VGGT-DD) are competitive but still struggle to produce accurate novel-view depths. This stems from the fact that they learn a direct mapping from observed depth to novel-view depth, where depth alone is not a sufficiently rich representation of full 3D scene geometry. In contrast, 3DFMs predict depth from a richer latent scene representation encoded in the backbone, with the depth head acting as a lightweight decoder. While this head captures some geometric cues, it is less expressive than the full latent features, which better support reasoning about scene structure from novel viewpoints. Z3D-VGGT and Z3D-WM consistently outperform our baselines in out-of-domain evaluations in point clouds accuracy and depth map metrics and perform on par with each other across all experiments. This indicates that the 3D representations learned by both 3DFMs capture similar geometric structures of the scene. We attribute the strong performance of Z3D-WM and Z3D-VGGT to the rich geometric scene representations learned by these models, as well as our implicit modeling of novel-depth synthesis through the diffusion framework.

Outdoor Evaluation. Our training/evaluation data includes outdoor data from MegaDepth. This was included in the ‘In Domain’ grouping; we explicitly pull out the MegaDepth results in Tab. 6. Z3D substantially outperforms the DD models in depth metrics, and has substantially higher point cloud accuracy with

Table 2: Point Cloud Metrics (1 source $\rightarrow$ 1 target). Best results per dataset in **bold**, second best underlined. Z3D performs well, especially in accuracy, and often has comparable completeness. Blue color indicates out of domain evaluation. Note: M means mean and Md means median

Model	In-Domain				DTU				Out of Domain 7-Scenes				NRGBD			
	Acc $\downarrow$		Comp $\downarrow$		Acc $\downarrow$		Comp $\downarrow$		Acc $\downarrow$		Comp $\downarrow$		Acc $\downarrow$		Comp $\downarrow$	
	M	Md	M	Md	M	Md	M	Md	M	Md	M	Md	M	Md	M	Md
LVSM+VGGT	-	-	-	-	7.681	6.503	23.396	11.958	0.065	0.042	0.184	0.091	0.168	0.097	0.369	<u>0.204</u>
LVSM+WM	-	-	-	-	7.590	6.304	23.242	11.512	0.063	0.038	0.178	0.083	0.168	0.095	0.385	0.217
VGGT-DD	0.154	0.044	0.149	0.013	6.337	4.386	1.644	1.155	0.046	0.029	<u>0.029</u>	<u>0.010</u>	0.255	0.092	0.192	0.743
WM-DD	0.130	0.053	0.068	0.013	11.282	7.379	<u>1.737</u>	<u>1.335</u>	0.042	0.029	0.024	0.009	<u>0.159</u>	0.084	<u>0.127</u>	0.837
Z3D-VGGT	<u>0.121</u>	0.020	0.181	<u>0.021</u>	2.725	1.782	2.938	2.225	<u>0.033</u>	<u>0.018</u>	0.046	0.022	0.340	<u>0.056</u>	0.260	0.676
Z3D-WM	0.083	<u>0.022</u>	<u>0.100</u>	<u>0.021</u>	<u>5.021</u>	<u>2.408</u>	3.827	2.610	0.029	0.017	0.038	0.019	0.154	0.050	0.133	0.826

Table 3: Depth Estimation Metrics (1 source $\rightarrow$ 1 target). Best results per dataset in **bold**, second best underlined. Blue color indicates out of domain evaluation.

Model	In-Domain		Out-of-Domain			
			DTU		7-Scenes	
	AbsRel $\downarrow$ $\delta < 1.25 \uparrow$		AbsRel $\downarrow$ $\delta < 1.25 \uparrow$		AbsRel $\downarrow$ $\delta < 1.25 \uparrow$	
LVSM+VGGT	-	-	0.468	0.024	0.342	0.745
LVSM+WM	-	-	0.468	0.025	0.328	0.749
VGGT-DD	0.098	0.914	0.022	<u>0.998</u>	0.274	<u>0.963</u>
WM-DD	0.108	0.902	0.036	0.997	0.272	0.967
Z3D-VGGT	<u>0.076</u>	<u>0.935</u>	0.012	0.999	0.260	<u>0.963</u>
Z3D-WM	0.075	0.939	<u>0.021</u>	0.999	<u>0.266</u>	0.967

slightly lower completeness than the DD. This also shows that Z3D method is domain agnostic and works for both indoor and outdoor scenes.

Generalization to Newer 3DFMs. We further evaluate Z3D on a recently introduced 3D foundation model, VGGT- Ω [46], demonstrating that the proposed framework is largely model-agnostic. We train Z3D-VGGT- Ω under the one-source-view $\rightarrow$ one-target-view setting without any algorithmic modifications. Tables 7 and 8 report the corresponding point cloud and depth estimation results. Z3D-VGGT- Ω achieves performance comparable to Z3D-VGGT and Z3D-WM, showing that Z3D can be readily adapted to new 3D foundation models while maintaining competitive performance.

Limitations. Z3D advances novel-view depth synthesis by leveraging 3DFM scene representations, but also has certain limitations. We observe that performance degrades when there is limited overlap between source and target views, making it difficult to recover fine-grained geometry in novel viewpoints. In our setup, the first source image is treated as the reference camera. Prior work on 3DFMs [50] has shown that the choice of reference view can affect model performance. Z3D inherits this sensitivity and reliable 3D reconstruction becomes challenging when input views provide insufficient scene coverage. Thus, with low-overlap conditions between source and target views, both Z3D-VGGT and Z3D-WM struggle to produce accurate depth predictions for novel views. Fig-

Table 4: Point Cloud Metrics (2 source $\rightarrow$ 4 target). **Bold:** best (lowest) per column per dataset, underline: second best.

Model	In-Domain				DTU				Out of Domain 7-Scenes				NRGBD			
	Acc $\downarrow$		Cmp $\downarrow$		Acc $\downarrow$		Cmp $\downarrow$		Acc $\downarrow$		Cmp $\downarrow$		Acc $\downarrow$		Cmp $\downarrow$	
	M	Md	M	Md	M	Md	M	Md	M	Md	M	Md	M	Md	M	Md
LVSM+VGGT	-	-	-	-	9.32	7.90	16.45	7.31	<u>0.025</u>	<u>0.014</u>	0.055	0.017	<u>0.062</u>	<u>0.028</u>	0.140	0.033
LVSM+WM	-	-	-	-	<u>9.20</u>	7.570	16.510	6.810	0.022	0.013	0.053	0.017	0.055	<u>0.028</u>	0.138	0.039
VGGT-DD	0.266	0.148	0.107	0.018	17.03	12.22	1.60	1.05	0.046	0.029	<u>0.023</u>	0.006	0.215	0.080	<u>0.091</u>	<u>0.018</u>
WM-DD	0.237	0.123	0.079	0.015	19.20	12.34	<u>1.690</u>	1.05	0.054	0.037	<u>0.023</u>	0.007	0.113	0.069	0.074	0.014
Z3D-VGGT	<u>0.131</u>	0.037	0.112	0.018	6.14	3.71	2.79	<u>1.96</u>	0.027	<u>0.014</u>	0.032	0.012	0.211	0.045	0.109	0.037
Z3D-WM	0.120	<u>0.041</u>	0.088	<u>0.019</u>	10.32	<u>5.67</u>	3.23	2.18	0.026	<u>0.014</u>	0.029	<u>0.010</u>	0.091	0.035	0.108	0.031

Table 5: Depth Estimation Metrics (2 sources $\rightarrow$ 4 targets). **Bold:** best per column per dataset, underline: second best.

Model	In-Domain		DTU		Out-of-Domain 7Scenes		NRGBD	
	AbsRel $\downarrow$ $\delta < 1.25 \uparrow$		AbsRel $\downarrow$ $\delta < 1.25 \uparrow$		AbsRel $\downarrow$ $\delta < 1.25 \uparrow$		AbsRel $\downarrow$ $\delta < 1.25 \uparrow$	
LVSM+VGGT	-	-	0.476	0.006	0.343	0.933	<u>0.127</u>	0.828
LVSM+WM	-	-	0.475	0.006	<u>0.346</u>	0.935	0.121	0.842
VGGT-DD	0.248	0.667	0.055	0.986	0.372	0.959	0.225	0.686
WM-DD	0.215	0.735	0.059	0.979	0.383	0.943	0.147	0.812
Z3D-VGGT	0.112	0.891	0.024	0.995	0.342	0.965	0.223	0.715
Z3D-WM	<u>0.118</u>	<u>0.889</u>	<u>0.036</u>	<u>0.987</u>	0.362	<u>0.962</u>	0.131	<u>0.830</u>

Table 6: MegaDepth depth, point cloud metrics. 2 source $\rightarrow$ 4 targets. **Bold:** best, underline: second best. M: Mean, Md: Median

Model	$\downarrow$		$\uparrow$		Acc $\downarrow$		Comp $\downarrow$	
	Abs Rel	$\delta < 1.25$	Mean	Med	Mean	Med	Mean	Med
VGGT-DD	0.102	0.905	0.203	0.058	<u>0.176</u>	<u>0.014</u>		
WM-DD	0.111	0.896	0.241	0.060	0.163	0.013		
Z3D-VGGT	0.039	<u>0.976</u>	0.091	0.020	0.213	0.023		
Z3D-WM	<u>0.041</u>	0.977	<u>0.110</u>	<u>0.025</u>	0.209	0.026		

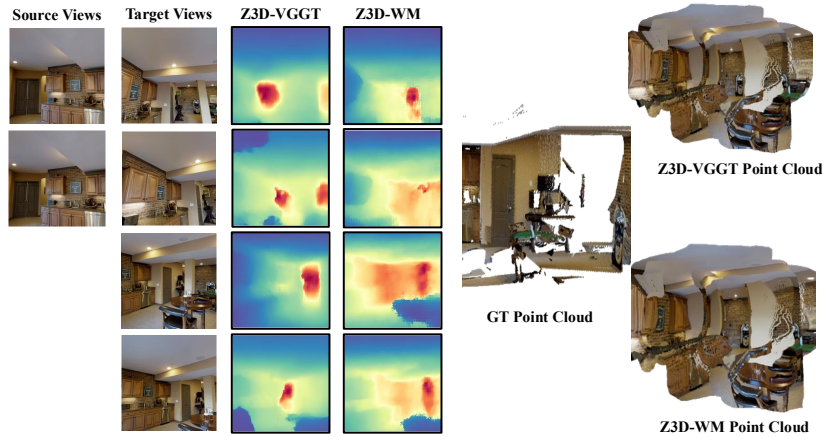
Table 7: Point Cloud Metrics (1 source $\rightarrow$ 1 target). **Bold:** best (lowest) per column per dataset, underline: second best.

Model	In-Domain				DTU				Out of Domain 7-Scenes				NRGBD			
	Acc $\downarrow$		Comp $\downarrow$		Acc $\downarrow$		Comp $\downarrow$		Acc $\downarrow$		Comp $\downarrow$		Acc $\downarrow$		Comp $\downarrow$	
	M	Md	M	Md	M	Md	M	Md	M	Md	M	Md	M	Md	M	Md
VGGT-DD	0.154	0.044	0.149	0.013	6.337	4.386	1.644	1.155	0.046	0.029	0.029	<u>0.010</u>	0.255	0.092	<u>0.097</u>	0.025
WM-DD	0.130	0.053	0.068	0.013	11.282	7.379	<u>1.737</u>	<u>1.335</u>	0.042	0.029	0.024	0.009	0.159	0.084	0.084	0.019
Z3D-VGGT	0.121	0.020	0.181	<u>0.021</u>	2.725	1.782	2.938	2.225	0.033	<u>0.018</u>	0.046	0.022	0.340	0.056	0.156	0.053
Z3D-WM	0.083	<u>0.022</u>	0.100	<u>0.021</u>	5.021	<u>2.408</u>	3.827	2.610	0.029	0.017	0.038	0.019	<u>0.154</u>	<u>0.050</u>	0.139	<u>0.050</u>
Z3D-VGGT- Ω	<u>0.108</u>	0.029	0.114	0.029	<u>4.834</u>	3.117	4.279	3.386	<u>0.030</u>	0.017	<u>0.032</u>	0.019	0.119	0.039	0.127	0.034

ure 7 illustrates a typical failure mode of Z3D, where the method struggles to predict novel-view depth with fine-grained details. In such cases, the resulting point clouds become inconsistent.

Table 8: Depth Estimation Metrics (1 sources $\rightarrow$ 1 target). **Bold:** best per column per dataset, underline: second best.

Model	In-Domain		DTU		Out-of-Domain 7-Scenes		NRGBD	
	AbsRel $\downarrow$	$\delta < 1.25 \uparrow$	AbsRel $\downarrow$	$\delta < 1.25 \uparrow$	AbsRel $\downarrow$	$\delta < 1.25 \uparrow$	AbsRel $\downarrow$	$\delta < 1.25 \uparrow$
VGGT-DD	0.098	0.914	0.022	0.998	0.274	0.963	0.192	0.743
WM-DD	0.108	0.902	0.036	0.997	<u>0.272</u>	<u>0.967</u>	<u>0.127</u>	0.837
Z3D-VGGT	<u>0.076</u>	<u>0.935</u>	0.012	0.999	0.260	0.963	0.260	0.676
Z3D-WM	0.075	0.939	0.021	0.999	0.266	<u>0.967</u>	0.133	0.826
Z3D-VGGT- Ω	0.096	0.925	<u>0.019</u>	0.999	0.305	0.978	0.126	0.837

**Fig. 7:** Point maps obtained from the depth maps predicted by Z3D-VGGT and Z3D-WM. Both Z3D-VGGT and Z3D-WM struggles to predict depths when source views have limited overlap with target views. The depth maps from Z3D-VGGT and Z3D-WM preserve the overall scene structure, but finer details are missing, and the resulting projected point clouds is not consistent.

5 Conclusion

We present Z3D, novel depth synthesis model that requires at atleast one view and a new view pose to predict the novel depth of that new view. Our method can take an arbitrary number of posed images and predict the depth of novel views conditioned on the camera pose and latent scene representation of the scene extracted from 3DFMs.

Acknowledgements

This work was partially supported by NSF #2437330 and NYU IT High Performance Computing resources, services, and staff expertise. Thanks to Sarah Jabbour, Joseph Tung, Ruoyu Wang, and V. Samuel Pérez-Díaz for helpful feedback.

References

1. Aanaes, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision* pp. 1–16 (2016)
2. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. *Commun. ACM* **54**(10), 105–112 (Oct 2011). <https://doi.org/10.1145/2001269.2001293>, <https://doi.org/10.1145/2001269.2001293>
3. Azinović, D., Martin-Brualla, R., Goldman, D.B., Nießner, M., Thies, J.: Neural rgb-d surface reconstruction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 6290–6301 (June 2022)
4. Banani, M.E., Raj, A., Maninis, K.K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., Jampani, V.: Probing the 3D Awareness of Visual Foundation Models. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 21795–21806. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2024). <https://doi.org/10.1109/CVPR52733.2024.02059>, <https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.02059>
5. Beck, A., Sabach, S.: Weiszfeld’s method: Old and new results. *J. Optim. Theory Appl.* **164**(1), 1–40 (Jan 2015). <https://doi.org/10.1007/s10957-014-0586-7>, <https://doi.org/10.1007/s10957-014-0586-7>
6. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: *Proceedings of the 41st International Conference on Machine Learning. ICML’24, JMLR.org* (2024)
7. Furukawa, Y., Hernández, C.: Multi-view stereo: A tutorial. *Found. Trends Comput. Graph. Vis.* **9**, 1–148 (2015), <https://api.semanticscholar.org/CorpusID:61831046>
8. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. *2007 IEEE Conference on Computer Vision and Pattern Recognition* pp. 1–8 (2007), <https://api.semanticscholar.org/CorpusID:2845053>
9. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (December 2015)
10. Guizilini, V., Irshad, M.Z., Chen, D., Shakhnarovich, G., Ambrus, R.: Zero-shot novel view and depth synthesis with multi-view geometric diffusion. In: *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. pp. 764–776 (June 2025)
11. Guizilini, V., Vasiljevic, I., Fang, J., Ambru, R., Shakhnarovich, G., Walter, M.R., Gaidon, A.: Depth field networks for generalizable multi-view scene representation. In: *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*. p. 245–262. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-19824-3_15, https://doi.org/10.1007/978-3-031-19824-3_15
12. Guizilini, V.C., Tokmakov, P., Dave, A., Ambrus, R.: Grin: Zero-shot metric depth with pixel-level diffusion. *2025 International Conference on 3D Vision (3DV)*. pp. 112–122 (2024), <https://api.semanticscholar.org/CorpusID:272689130>
13. Guizilini, V.C., Vasiljevic, I., Chen, D., Ambrus, R., Gaidon, A.: Towards zero-shot scale-aware monocular depth estimation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* pp. 9199–9209 (2023), <https://api.semanticscholar.org/CorpusID:259309440>

14. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, USA, 2 edn. (2003)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
16. Jang, W., Weinzaepfel, P., Leroy, V., Agapito, L., Revaud, J.: Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1071–1081 (2025), <https://api.semanticscholar.org/CorpusID:277244721>
17. Jiang, H., Tan, H., Wang, P., Jin, H., Zhao, Y., Bi, S., Zhang, K., Luan, F., Sunkavalli, K., Huang, Q., Pavlakos, G.: Rayzer: A self-supervised large view synthesis model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4918–4929 (October 2025)
18. Jin, H., Jiang, H., Tan, H., Zhang, K., Bi, S., Zhang, T., Luan, F., Snively, N., Xu, Z.: Lvsm: A large view synthesis model with minimal 3d inductive bias. In: The Thirteenth International Conference on Learning Representations (2025), <https://openreview.net/forum?id=QQBPWtvtcn>
19. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
20. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4) (Jul 2023). <https://doi.org/10.1145/3592433>, <https://doi.org/10.1145/3592433>
21. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r (2024)
22. Li, Z., Snively, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
23. Liu, Y., Min, Z., Wang, Z., Wu, J., Wang, T., Yuan, Y., Luo, Y., Guo, C.: World-mirror: Universal 3d world reconstruction with any-prior prompting. arXiv preprint arXiv:2510.10726 (2025)
24. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (Dec 2021). <https://doi.org/10.1145/3503250>, <https://doi.org/10.1145/3503250>
25. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (Jul 2022). <https://doi.org/10.1145/3528223.3530127>, <https://doi.org/10.1145/3528223.3530127>
26. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S.M., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5470–5480 (2021), <https://api.semanticscholar.org/CorpusID:244773517>
27. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6309–6318. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)

28. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4172–4182 (2023). <https://doi.org/10.1109/ICCV51070.2023.00387>
29. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 1623–1637 (2019), <https://api.semanticscholar.org/CorpusID:195776274>
30. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12159–12168 (2021). <https://doi.org/10.1109/ICCV48922.2021.01196>
31. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 10912–10922 (October 2021)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
33. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: *ACM SIGGRAPH 2022 Conference Proceedings. SIGGRAPH '22*, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3528233.3530757>, <https://doi.org/10.1145/3528233.3530757>
34. Saharia, C., Chan, W., Saxena, S., Lit, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22*, Curran Associates Inc., Red Hook, NY, USA (2022)
35. Sajjadi, M.S.M., Mahendran, A., Kipf, T., Pot, E., Duckworth, D., Lučić, M., Greff, K.: Rust: Latent neural scene representations from unposed imagery. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17297–17306 (2023). <https://doi.org/10.1109/CVPR52729.2023.01659>
36. Sajjadi, M.S.M., Meyer, H., Pot, E., Bergmann, U., Greff, K., Radwan, N., Vora, S., Lucic, M., Duckworth, D., Dosovitskiy, A., Uszkoreit, J., Funkhouser, T., Tagliasacchi, A.: Scene Representation Transformer: Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations. *CVPR* (2022), <https://srt-paper.github.io/>
37. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9338–9346. IEEE Computer Society, Los Alamitos, CA, USA (Nov 2019). <https://doi.org/10.1109/ICCV.2019.00943>, <https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00943>
38. Saxena, S., Herrmann, C., Hur, J., Kar, A., Norouzi, M., Sun, D., Fleet, D.J.: The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23*, Curran Associates Inc., Red Hook, NY, USA (2023)
39. Saxena, S., Hur, J., Herrmann, C., Sun, D., Fleet, D.J.: Zero-shot metric depth with a field-of-view conditioned diffusion model (2023)

40. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113 (2016). <https://doi.org/10.1109/CVPR.2016.445>
41. Shade, J., Gortler, S.J., Wei He, L., Szeliski, R.: Layered depth images. Proceedings of the 25th annual conference on Computer graphics and interactive techniques (1998), <https://api.semanticscholar.org/CorpusID:1240104>
42. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in rgb-d images. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. p. 2930–2937. CVPR '13, IEEE Computer Society, USA (2013). <https://doi.org/10.1109/CVPR.2013.377>, <https://doi.org/10.1109/CVPR.2013.377>
43. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H.M., Nardi, R.D., Goesele, M., Lovegrove, S., Newcombe, R.: The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
44. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. arXiv preprint arXiv:2408.16061 (2024)
45. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupperecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
46. Wang, J., Chen, M., Zhang, S., Karaev, N., Schönberger, J., Labatut, P., Bojanowski, P., Novotny, D., Vedaldi, A., Rupperecht, C.: VGGT- Ω . In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
47. Wang, J., Rupperecht, C., Novotny, D.: Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9739–9749 (2023). <https://doi.org/10.1109/ICCV51070.2023.00896>
48. Wang*, Q., Zhang*, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: CVPR (2025)
49. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024)
50. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3 : Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
51. Xu, G., Lin, H., Luo, H., Wang, X., Yao, J., Zhu, L., Pu, Y., Chi, C., Sun, H., Wang, B., et al.: Pixel-perfect depth with semantics-prompted diffusion transformers. arXiv preprint arXiv:2510.07316 (2025)
52. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2025)
53. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to Recover 3D Scene Shape from a Single Image. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 204–213. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2021). <https://doi.org/10.1109/CVPR46437.2021.00027>, <https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00027>

54. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural radiance fields from one or few images. In: CVPR (2021)
55. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
56. Zhang, J.Y., Lin, A., Kumar, M., Yang, T.H., Ramanan, D., Tulsiani, S.: Cameras as rays: Pose estimation via ray diffusion. In: International Conference on Learning Representations (ICLR) (2024)
57. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 3813–3824 (2023), <https://api.semanticscholar.org/CorpusID:256827727>
58. Zhang, Y., Tung, J., Cai, R., Fouhey, D., Averbuch-Elor, H.: Emergent extreme-view geometry in 3d foundation models (2025), <https://arxiv.org/abs/2511.22686>
59. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. ICCV (2023)
60. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders (2025)