Matrix3D: Large Photogrammetry Model All-in-One

Yuanxun Lu[1][*][†]   Jingyang Zhang[2][*]   Tian Fang[2]   Jean-Daniel Nahmias[2]   Yanghai Tsin[2]
Long Quan[3]   Xun Cao[1]   Yao Yao[1][‡]   Shiwei Li[2]

[1]Nanjing University

luyuanxun@smail.nju.edu.cn, {caoxun,yaoyao}@nju.edu.cn

[2]Apple

{jingyang_zhang,fangtian,jnahmias,ytsin,shiwei}@apple.com

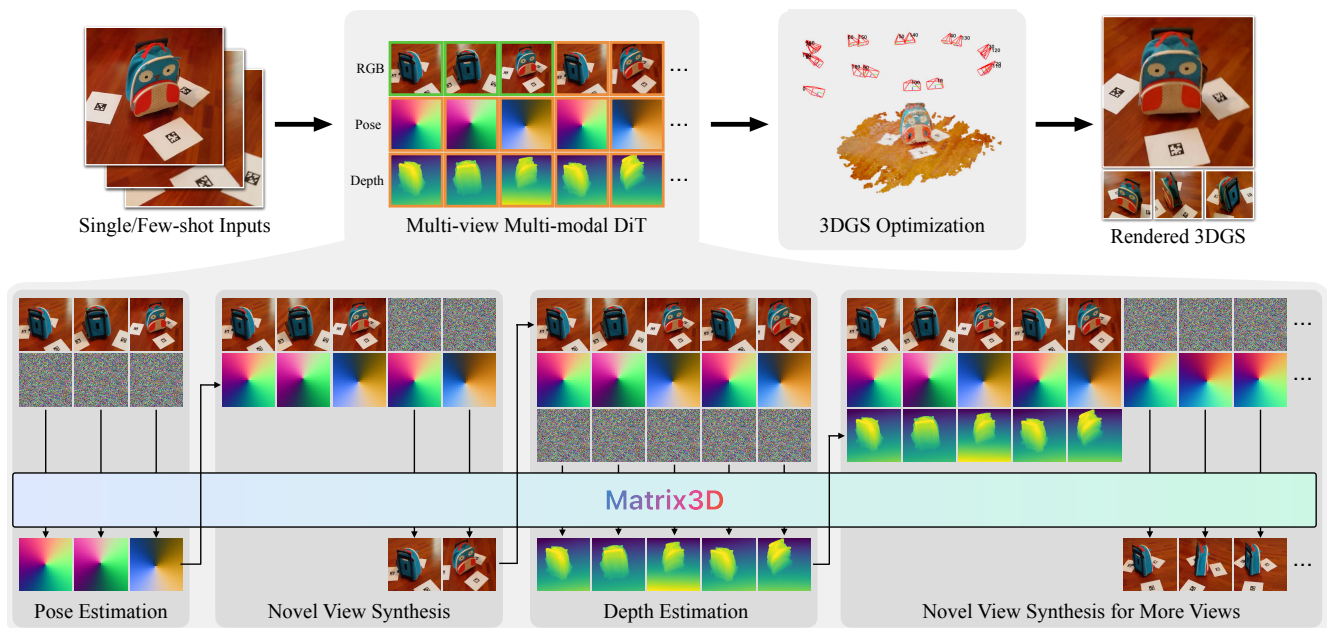[3]The Hong Kong University of Science and Technology

quan@cse.ust.hk

Figure 1. Utilizing **Matrix3D** for single/few-shot reconstruction. Before 3DGS optimization, we complete the input set by pose estimation, depth estimation and novel view synthesis, all of which are done by **the same** model.

Abstract

We present Matrix3D, a unified model that performs several photogrammetry subtasks, including pose estimation, depth prediction, and novel view synthesis using just the same model. Matrix3D utilizes a multi-modal diffusion transformer (DiT) to integrate transformations across several modalities, such as images, camera parameters, and depth maps. The key to Matrix3D's large-scale multi-modal training lies in the incorporation of a mask learning strategy. This enables full-modality model training even with partially complete data, such as bi-modality data of image-pose and image-depth pairs, thus significantly increases the pool of available training data. Matrix3D demonstrates state-of-the-art performance in pose estimation and novel view synthesis tasks. Additionally, it offers fine-grained control through multi-round interactions, making it an innovative tool for 3D content creation. Project page: https://nju-3dv.github.io/projects/matrix3d.

---

[*]Equal contribution.

[†]This project was performed during Yuanxun Lu's internship at Apple.

[‡]Corresponding author.

1. Introduction

Photogrammetry is a crucial technology for reconstructing 3D scenes from 2D images. However, the traditional photogrammetry pipeline has two significant weaknesses. First, it typically requires a dense collection of images–often hundreds–to achieve robust and accurate 3D reconstruction, which can be troublesome in practical applications. Second, the pipeline involves multiple processing stages that utilize completely different algorithm blocks, including feature detection, structure-from-motion (SfM), multi-view stereo (MVS), and etc. While each step is an independent task that requires unique algorithm, they are not correlated or jointly optimized with one another, which can result in suboptimal outcomes and accumulated errors throughout this multi-stage process.

The first challenge is commonly tackled by combining reconstruction with generation [29, 49, 92, 115], where denser RGB images are generated by diffusion models conditioned on sparse inputs. In practice, however, when there are more than one input image, it is challenging to obtain accurate relative poses of them since they often come with low overlaps. The second challenge has not yet been extensively addressed. Representative works are PF-LRM [110] and DUSt3R [111], which use single feed-forward models to perform both pose estimation and scene reconstruction. They are thus end-to-end optimizable and eliminate the need of multi-stage processing.

Inspired by previous methods, we try to take one step further and tackle these two challenges together, by building a unified model that can do multiple photogrammetry subtasks, including pose estimation, depth estimation and novel view synthesis (for sparse view reconstruction). We call this model **Matrix3D**, featuring an all-in-one generative model designed to support various sub-tasks in photogrammetry, through altering input/output combinations. At its core, Matrix3D represents data of all modalities using unified 2D representations: camera geometries are encoded as Plücker ray maps, while 3D structures are presented as 2.5D depth maps. This makes it possible to leverage the capabilities of modern image generative models. We extend the diffusion transformer into a multi-view, multi-modal framework, capable of generating all necessary modalities. Inspired by the principles of masked auto-encoder (MAE), our model is trained by randomly masking inputs, while predicting the remaining unseen observations. This masking learning design not only effectively manages varying degrees of input sparsity, but also substantially increases the volume of available training data by utilizing partially complete data samples such as bi-modality image-pose and image-depth pairs.

With the densified camera / image / depth predictions generated by Matrix3D, a 3D Gaussian Splatting (3DGS) [45] optimization can be applied to produce the final output, where depth maps are back-projected into 3D point clouds for 3DGS initialization.

In summary, our key contribution is the unified diffusion transformer model that has flexible input/output configurations and enables several tasks in photogrammetry process. The proposed method eliminates the need for multiple task-specific models, streamlining the photogrammetry process with one single model. Extensive quantitative and qualitative experiments, demonstrate the state-of-the-art performance of our methods compared with existing task-specific approaches.

2. Related Work

Photogrammetry, also known as image-based 3D reconstruction, is a foundational pillar in the field of 3D vision. A typical photogrammetry pipeline consists of several critical sub-steps. Below, we provide an overview of the most relevant works related to image-based 3D reconstruction.

Structure-from-Motion (SfM) is a classical approach for simultaneously recovering sparse 3D structures and estimating camera poses from multiple overlapping 2D images. These pipelines [17, 18, 35, 87] typically begin with camera parameter estimation through feature matching [5, 34, 63, 83] across images, followed by bundle adjustment to jointly optimize the 3D point cloud and camera poses. Recent advances have focused on enhancing the robustness of SfM through learning-based feature extractors [21, 24, 65, 80], improved image-matching techniques [12, 56, 86, 96], and neural bundle adjustment [53, 55, 108, 117]. However, challenges still persist in sparse input scenarios, though gradually alleviated. Limited observations still introduce multi-view ambiguity and performance degradation.

Multi-view Stereo (MVS) builds upon the camera poses obtained by SfM to create dense 3D geometry. Traditional MVS methods [26, 27, 37, 88] depend on hand-crafted features and engineered regularizations to build dense correspondences and recover 3D points [25, 48], volumes [25, 46, 90], or depth maps [8, 27, 102]. Learning-based methods [14, 31, 127, 128, 134] offer more powerful reconstruction with improved completeness and generalization capabilities. Typically, these methods assume well-calibrated camera parameters, which limits their robustness and applicability in real-world scenarios where calibration may be imprecise or unavailable.

Sparse-view Pose Estimation is extremely challenging due to the limited overlap between input images, making it difficult for traditional methods [69, 94] to build correspondences. Recent research has explored various strategies to address these ambiguities and predict relative poses from sparsely sampled images, including energy optimization [52, 93], exploiting synthetic data [43], leveraging data priors [107], and applying probabilistic models [13, 136]. Recently, RayDiffusion [137] employs ray-based cameras and diffusion methods to model pose distributions, while
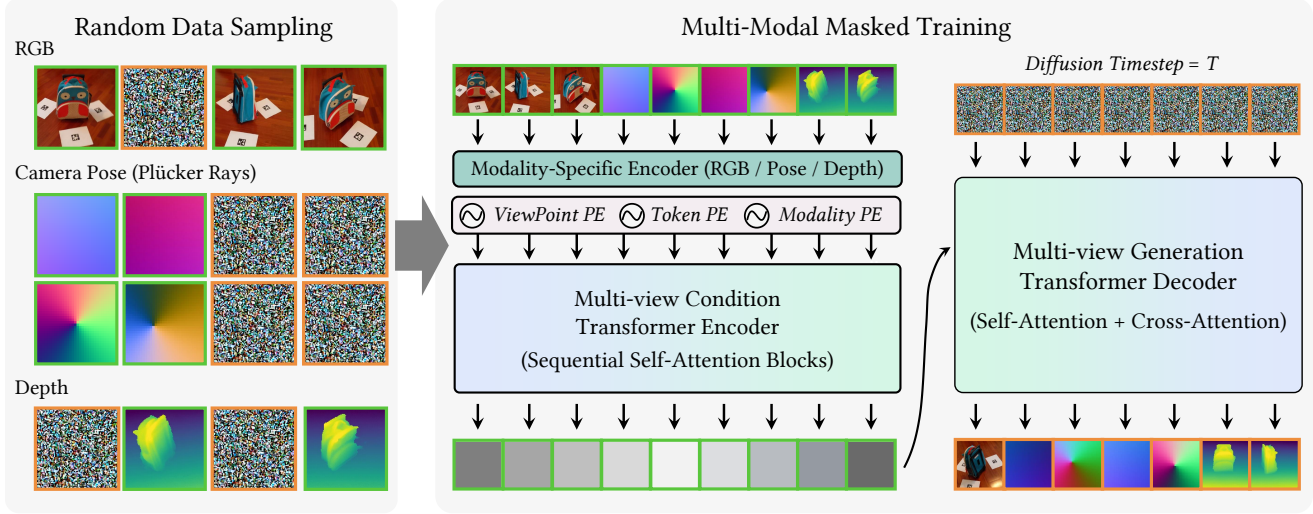
Figure 2. We train the Matrix3D by masked learning. Multi-modal data are randomly masked by noise corruption. Observations (green) and noisy maps (yellow) are fed into the encoder and the decoder respectively. By attaching the view and modality information to the clean and noisy inputs via different positional encodings, the model learns to denoise the corrupted maps and generate the desired outputs.

PF-LRM [110] and DUSt3R [111] predict point maps in reference frames and recover poses using PnP algorithms.

Feed-forward RGB-to-3D approaches aim to directly infer 3D representations from single or a few RGB images via feed-forward models, without requiring per-scene optimization. These methods leverage strong 3D priors learned from large-scale object-level [11] or larger datasets [19] to address inherent ambiguities. Pre-trained geometry models, such as monocular depth predictors [6, 77, 78, 130], are usually adopted to further improve robustness. Recently, people proposed feed-forward methods [40, 49, 99, 116, 121, 138] which combines large transformer models to directly map RGB images into 3D representations [9, 28, 54]. While these methods enable efficient 3D generation, their results remain less accurate than optimization-based methods.

3D Generation with 2D Priors refers to methods that use pre-trained 2D vision models to guide 3D generation. DreamFusion [73] first proposed Score Distillation Sampling (SDS) to synthesize NeRFs from text by iteratively distilling knowledge from text-to-image diffusion models. Subsequent research has enhanced performance by improving distillation strategies [15, 33, 42, 98, 112, 133, 135] or fine-tuning 2D diffusion models with camera conditioning [10, 30, 58, 59, 76, 85, 115]. Recent studies have shown that fine-tuning the model to generate multi-view images simultaneously [57, 91, 92, 100, 123] provides stronger priors. Beyond SDS, researchers also explored offline 3D reconstruction [29, 60, 61, 64] directly from the generated multi-view information. Moreover, fine-tuning video diffusion models, which inherently encode 3D knowledge, has emerged as a promising approach [32, 47, 67, 104], though their higher computational demands remain a challenge.

Masked Learning has achieved significant success in pre-training tasks for NLP [7, 44, 74, 75] and computer vision [2, 36, 50]. These methods are proven to capture high-level semantics by masking parts of input and training models to reconstruct the masked contents. Recent research has extended this idea to multi-view image settings [113, 114], demonstrating improvements in downstream tasks like optical flow and stereo matching. In this work, we further apply masked learning to multi-view and multi-modal training to develop an all-in-one photogrammetry model.

3. Method

In this section, we introduce Matrix3D, an all-in-one photogrammetry model for unified 3D reconstruction and generation. In the following, we describe the details of the framework design (Section 3.1), masking strategies (Section 3.2), dataset preparation (Section 3.3), training setup (Section 3.4), and downstream tasks (Section 3.5).

3.1. Multi-Modal Diffusion Transformer

As demonstrated in Section 1, our framework is designed around three key principles: a unified probabilistic model, flexible I/O, and multi-task capability. The emerging diffusion transformer (DiT) [71] offers an ideal foundation, with its transformer architecture naturally supporting flexible I/O configurations and multi-modal fusion.

Network architecture. The proposed model consists of two novel components compared with a standard image DiT model: a multi-view encoder and a multi-view decoder D. The encoder processes conditioning data from multiple views across different modalities (i.e., RGB, poses and depth), and embeds them into a shared latent space, enabling better cross-view and cross-modal feature integration. Similarly, the decoder processes noisy maps corre-

sponding to different targets at different diffusion timestamps. The latent codes from different views and modalities are concatenated sequentially and passed through transformer layers to capture correspondence across views and modalities. Both the encoder and decoder are composed of multiple self-attention blocks, with the decoder additionally incorporating a cross-attention block after each self-attention layer to enable communication between conditioning inputs and generated outputs. Our diffusion model is built upon the pre-trained Hunyuan-DiT [51] architecture with the aforementioned module modifications. In our experiments, the maximum number of views is set to 8, though this number can be further extended subject to computation budget. Mathematically, let $\mathbf{x}_c$ denotes multi-view/modality conditions, $\mathbf{x}_{g,t}$ the desired generation corrupted by noise ϵ at time t, and $\mathbf{x}_0$ the groundtruth. The diffusion model is trained using v-prediction [84] loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, \epsilon, t, y} \left[\| D(E(\mathbf{x}_c), \mathbf{x}_{g,t}, t) - \mathbf{v} \|^2 \right], \quad (1)$$

$$\mathbf{v} = \alpha_t \epsilon - \sigma_t \mathbf{x}_0. \quad (2)$$

Multi-modality encoding. To handle multiple modalities, we apply modality-specific encoding methods before feeding the data into transformers. Specifically, the VAE from SDXL [72] is used to encode RGB images into low-dimensional latent space. For camera poses, we follow Ray-Diffusion [137] to represent cameras as Plücker ray maps, which naturally takes the form of image-like 2D data. For depth modality, our model adopts multi-view aligned depth (i.e., affine-invariant depth), which would be converted into disparities (i.e., the inverse of depth) to ensure a more compact data range. Additionally, a fixed shift and scale factor is applied to regularize ray maps and depth maps so that their distribution is closer to standard Gaussian distribution, as required by the diffusion process [39].

Positional encoding. We incorporate three types of positional encodings to preserve spatial relationships across viewpoints, patch token positions, and modalities. Specifically, we apply Rotary Positional Embedding (RoPE) [95] to encode the positions of individual patch tokens because we care more about their relative position, while apply absolute sinusoidal positional encoding [22] to viewpoints and modalities, each with different base frequencies, because we only need to distinguish between different view or modality IDs.

3.2. Masked Learning

Unlike masked autoencoders (MAE), which mask portions of a single image, we extend this concept to the image level across multi-view, multi-modal settings, following a similar approach to 4M [68], to enable flexible I/O configurations. By masking specific views or modalities, the model learns

to predict the missing content during both training and inference, facilitating dynamic and adaptable task handling.

Training strategy. During training, in addition to the standard fully random masking strategy, we apply task-specific assignments. Specifically, we divide the training tasks into novel view synthesis, pose estimation, and depth prediction, along with the full random tasks, following a 3:3:3:1 ratio. We adopt a multi-stage training strategy: first training on 4-view models at 256 resolution, followed by 8-view models, and finally on 8-view models at 512 resolution. Following Hunyuan-DiT [51], we use v-prediction [84] as training objective. We adopt a 10% probability of dropping conditions to enable classifier-free guidance (cfg) [38].

3.3. Dataset Preparation

Training data. We train Matrix3D on a mixture of six datasets: Objaverse [19], MVImgNet [132], CO3D-v2 [79], RealEstate10k [141], Hypersim [81], and ARKitScenes [4]. The first three datasets are indoor and object-centric, while the latter three cover large-scale indoor and outdoor scenes. Note that not all datasets provide all modalities and we use the available modalities for each dataset accordingly.

Normalization. Due to the highly diverse distributions of existing datasets, including variations in scale and scene type, preprocessing them consistently poses a challenge. To address this, we apply scene normalization and camera normalization. Please check the supplementary for details.

Incomplete depth. Real-world datasets often provide incomplete depth. A default solution to use these data is discarding patches with invalid pixels, which results in low data utilization. Instead, we concatenate valid masks to all the depth maps for both condition maps and noisy maps, and let the model learn how to discard invalid pixels. This also allow us to utilize sparse depth input during inference.

3.4. Training Setup

We initialize our model with pre-trained Hunyuan-DiT checkpoints [51]. In the first stage, we train the model for 180K steps with a learning rate of 1e-4 on 64 NVIDIA A100 80G GPUs. In the latter two stages, we train the model for 20K steps for each stage with a learning rate of 1e-5 on 128 GPUs. The entire training takes approximately 20 days.

3.5. Downstream Tasks

Reconstruction tasks as modality conversions. The Matrix3D model is able to perform multiple downstream tasks including pose estimation, multi-view depth estimation, novel view synthesis, or any hybrid of these tasks. By feeding any combination of conditional information and the desired output maps as noise, the model denoises the outputs according to the learned joint distribution. For example, pose estimation can be conducted by providing images of all
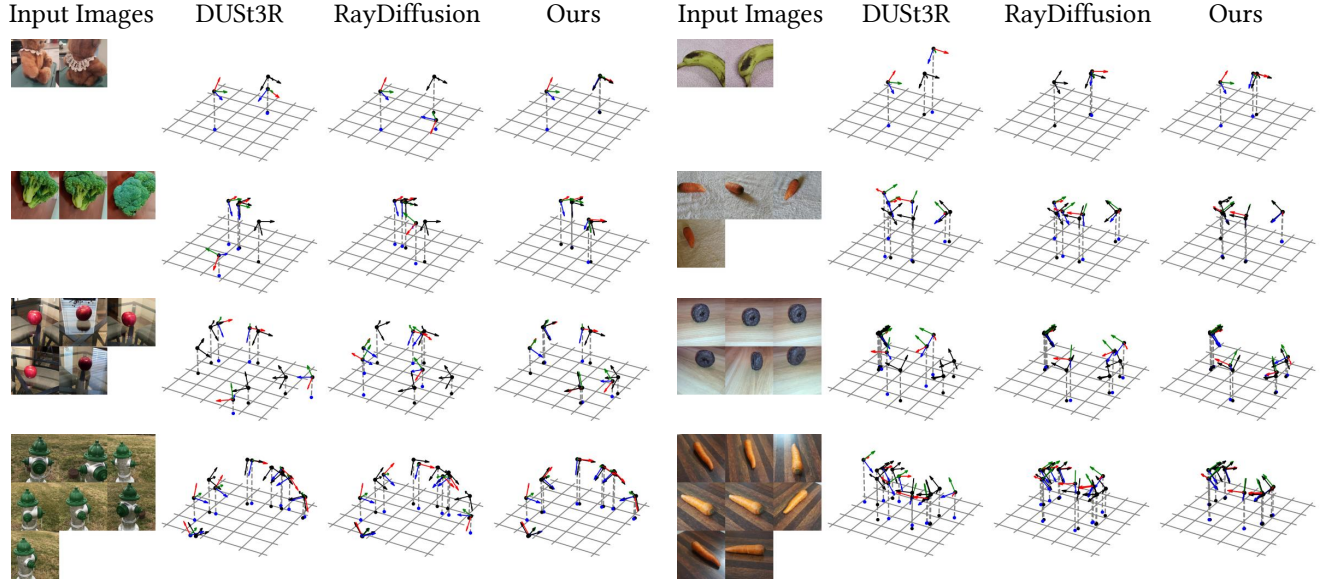
Figure 3. Sparse-view pose estimation results on CO3D dataset. The black axes are ground-truth and the colored ones are the estimation.
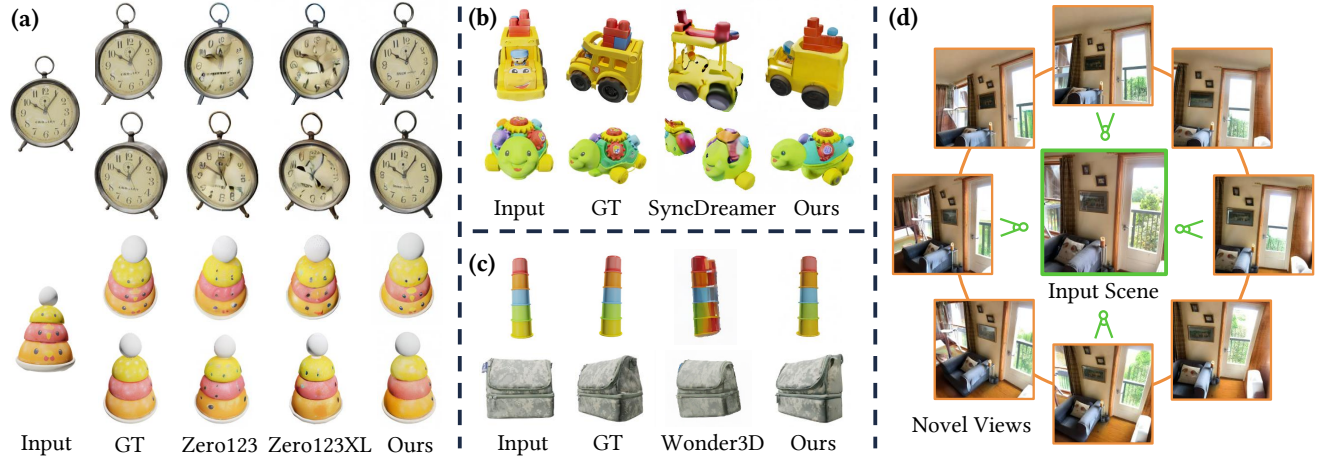

Figure 4. Qualitive evaluation results of novel view synthesis from single images on GSO and ARKitScenes dataset: a) random novel views; b) and c) follow the view configuration of SyncDreamer and Wonder3D respectively; d) indoor scenes from ARKitScenes dataset. Note that our method supports NVS of **arbitrary poses**.

input views, the identity camera pose for the reference view, and noisy ray maps for other views; novel view synthesis can be formulated as providing posed images for all reference views, ray maps for the novel views, and noisy novel-view images. Moreover, our model allows any reasonable input/output combinations, which cannot be achieved by previous task-specific models. For example, Matrix3D can achieve better results during NVS and pose estimation if depth maps are additionally provided.

3DGS Optimization. In this section, we describe how to utilize our model for single or few-shot image reconstruction. We first use the Matrix3D model to complete multimodality input (images/poses/depth maps) and also the image viewpoints. For few-shot inputs, we 1) estimate their camera poses, which can only be achieved by external methods in previous few-shot reconstruction systems; 2) estimate depth for the inputs as an initialization for 3DGS optimization; 3) synthesize novel views for stabilizing 3DGS optimization. For single input, we 1) synthesize more images to reach 8 key views with relatively large baselines to have an overall coverage of the target object or scene; 2) estimate depth maps for these key views; 3) synthesize novel views to interpolate the key images. Finally, we forward all the previous results to an open-source 3DGS reconstruction [97] with several modifications. It is noteworthy that the 3DGS reconstruction is specially tailored to mitigate the multi-view inconsistency among the generated images.

Please check the supplementary materials for more details about architectures, data pre-processing, training, and 3DGS optimization designs.

Metrics	Relative Rotation Accuracy @ 15°(↑, %)							Camera Center Accuracy @ 0.1 (↑, %)						
# of Images	2	3	4	5	6	7	8	2	3	4	5	6	7	8
COLMAP (SP+SG) [87]	31.3	29.0	27.3	27.6	28.0	29.3	31.9	100.0	34.8	24.1	19.1	16.0	15.0	15.7
PoseDiffusion [107]	73.6	74.3	74.6	75.4	76.0	76.7	76.9	100.0	75.1	66.4	62.5	60.2	59.1	58.1
RelPose++ [52]	79.8	80.8	82.0	82.7	83.0	83.4	83.7	100.0	82.5	74.7	70.7	68.2	66.5	65.0
DUSt3R [111]	85.6	88.6	90.1	90.7	91.3	91.7	92.0	100.0	87.8	83.9	81.3	80.3	80.1	79.2
RayDiffusion [137]	90.4	91.2	91.5	91.9	92.1	92.3	92.4	100.0	93.1	88.9	86.0	84.1	82.8	81.9
Ours RGB Only	<u>95.6</u>	<u>96.0</u>	**96.3**	**96.5**	**96.5**	<u>96.3</u>	**96.1**	100.0	<u>93.5</u>	<u>91.7</u>	<u>90.6</u>	<u>90.0</u>	<u>89.1</u>	<u>87.8</u>
Ours RGB + Depth	**95.8**	**96.3**	<u>96.2</u>	**96.5**	**96.5**	**96.4**	**96.3**	100.0	**93.8**	**92.0**	**91.5**	**91.0**	**90.4**	**89.5**

Table 1. Pose evaluation on CO3D. The percentage of relative rotations within 15 degrees and camera center errors within 10% of the groundtruth scene scale are reported. The best results are in **bold** and the second bests are <u>underlined</u>.

View Settings	Methods	PSNR ↑	SSIM ↑	LPIPS ↓
$\phi \in \{0°, 360°\}$ $\theta = 30°$	SyncDreamer [60]	<u>19.22</u>	<u>0.82</u>	**0.16**
	Ours	**20.45**	**0.86**	**0.16**
$\phi \in \{0°, 360°\}$ $\theta = 0°$	Wonder3D [61]	<u>13.28</u>	<u>0.78</u>	<u>0.33</u>
	Ours	**18.97**	**0.86**	**0.18**
$\phi \in \{0°, 360°\}$ $\theta \in \{30°, -20°\}$	InstantMesh [119]	13.78	0.80	0.25
	Ours	**18.66**	**0.85**	**0.19**
$\phi \in \{0°, 360°\}$ $\theta \in \{-90°, 90°\}$	Zero123 [59]	17.56	0.80	<u>0.18</u>
	Zero123-XL [20]	18.75	0.81	**0.17**
	Ours	<u>18.87</u>	<u>0.85</u>	0.21
	Ours +Depth	**19.16**	**0.86**	0.19

Table 2. Novel view synthesis (diffusion samples) evaluation on GSO. Results are grouped by different view settings because some methods uses fixed poses. Best results are in **bold** and the second best are <u>underlined</u>. ϕ and θ denote azimuth and elevation angles.

Method	δ_1↑	δ_2↑	δ_3↑	AbsRel↓	log10↓	RMS↓
Metric3D v2 [41, 131]	0.969	0.992	0.996	0.064	0.039	75.538
Depth Anything v2 [124, 125]	0.950	0.992	0.997	0.077	0.045	85.188
Ours CFG=1.5	0.985	0.997	0.999	**0.036**	0.023	47.806
Ours w/o CFG	**0.992**	**0.999**	**1.000**	0.038	**0.022**	**40.214**

Table 3. Quantitative evaluation of monocular metric depth prediction tasks on DTU. The best results are in **bold**.

4. Experiments

In the following, we present the experiment results of different photogrammetry tasks.

4.1. Pose Estimation

We first evaluate our model for pose estimation under sparse views on the CO3D dataset. The proposed model is compared with multiple types of previous works: 1) traditional SfM: COLMAP [87]; 2) discriminative neural networks: RelPose++ [52] and DUSt3R [111]; 3) generative neural networks: PoseDiffusion [107] and RayDiffusion [137].

We evaluate two metrics: relative rotation accuracy and camera center accuracy following RayDiffusion. For each scene, we estimate all 7 source views out of 8 views in one batch. The metrics are evaluated in a pair-wise manner. Our method outperforms other baselines by a significant margin (Table 1). Figure 3 also presents a qualitative comparison between our predictions and ground truth poses. Our method performs better by a large margin than all baselines.

4.2. Novel View Synthesis

In this section, we benchmark novel view synthesis task on diffusion samples on GSO [23] dataset against prior multi-view diffusion methods, including Zero123 [59], Zero123XL [20], SyncDreamer [60], Wonder3D [61], and InstantMesh [119]. For methods generating novel views at fixed camera poses, we follow their view configuration and use different rendered ground truths when comparing with them. For methods allowing arbitrary view synthesis, we render 32 random viewpoints for evaluation. PSNR, SSIM, and LPIPS are adopted as evaluation metrics. Figure 4 shows qualitative results on CO3D and ARKitScenes dataset. The quantitative results are shown in Table 2. Our method achieves the best results for most of the metrics.

4.3. Depth Prediction

Monocular Depth Prediction. We first evaluate our model on the monocular metric depth prediction task. Although our model is trained on at least two views, we found that it can still predict high-quality depth from single images. Table 3 shows the comparison against Metric3D v2 [41] and Depth Anything v2 [125]. We adopt the IDR [129] subset of the DTU dataset, and metrics following previous methods [41] to evaluate all methods. All methods are not trained on the DTU dataset. Our method performs significantly better than the baselines. Qualitatively, we found that monodepth methods often produce distorted geometry which cannot be recovered by global linear alignment. This may be resulted from focal ambiguity issue and large domain gap between object-centric and open-scene data.

Multi-view Depth Prediction. We then evaluate Matrix3D on the multi-view stereo depth prediction task. We evaluate it on the DTU [1] dataset. Following DUSt3R, we compute the Absolute Relative Error (rel) and Inlier Ratio on all test sets. Table 4 shows the quantitative results. We also back-project the depth maps to point clouds (with and without pose input) and evaluate their Chamfer distance to the ground truths. Quantitative results on DTU are shown in Table 5. More details about the conversion from depth maps to point clouds can be found in the supplementary material.

	Methods	Pose	Range	Int.	Align	rel↓	τ↑
(a)	COLMAP [87, 88]	✓	×	✓	×	**0.7**	**96.5**
	COLMAP Dense [87, 88]	✓	×	✓	×	20.8	69.3
(b)	MVSNet [127]	✓	✓	✓	×	(1.8)	(86.0)
	MVSNet Inv. Depth [127]	✓	✓	✓	×	(1.8)	(86.7)
	Vis-MVSNet [134]	✓	✓	✓	×	(1.8)	(87.4)
	MVS2D DTU [126]	✓	✓	✓	×	(3.6)	(64.2)
(c)	DeMon [103]	×	×	✓	$\|t\|$	21.8	16.6
	DeepV2D [101]	×	×	✓	med	7.7	33.0
	DUSt3R 224 [111]	×	×	×	med	**2.76**	**77.32**
	DUSt3R 512 [111]	×	×	×	med	3.52	69.33
(d)	DeMon [103]	✓	×	✓	×	23.7	11.5
	Deepv2D [101]	✓	×	✓	×	9.2	27.4
	MVSNet [127]	✓	×	✓	×	(4429.1)	(0.1)
	MVSNet Inv. Depth [127]	✓	×	✓	×	(28.7)	(48.9)
	Vis-MVSNet [134]	✓	×	✓	×	(374.2)	(1.7)
	MVS2D DTU [126]	✓	×	✓	×	(1.6)	(92.3)
	Robust MVD Baseline [89]	✓	×	✓	×	2.7	82.0
	Ours	✓	×	✓	×	**1.83**	**85.45**

Table 4. Quantitative evaluation of multi-view metric depth prediction tasks on DTU. The methods are categorized into: a) traditional methods; b) with poses and depth range; c) without poses and depth range, but with alignment; and d) with poses, without range and alignment. Parentheses refers to training on the same set. The best results are in **bold**. Results of other methods are reported in DUSt3R [111].
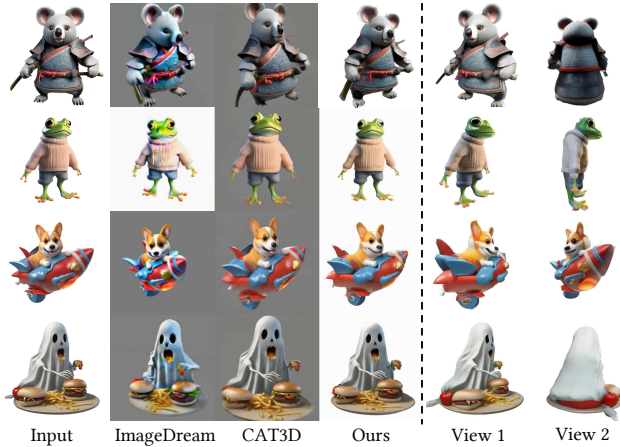
	Methods	GT cams	Acc.↓	Comp.↓	Overall↓
(a)	Camp [8]	✓	0.835	0.554	0.695
	Furu [25]	✓	0.613	0.941	0.777
	Tola [102]	✓	0.342	1.190	0.766
	Gipuma [27]	✓	**0.283**	0.873	0.578
(b)	MVSNet [127]	✓	0.396	0.527	0.462
	CVP-MVSNet [122]	✓	0.296	0.406	0.351
	UCS-Net [16]	✓	0.338	0.349	0.344
	CIDER [120]	✓	0.417	0.437	0.427
	CasMVSNet [31]	✓	0.325	0.385	0.355
	PatchmatchNet [105]	✓	0.427	0.277	0.352
	Vis-MVSNet [134]	✓	0.369	0.361	0.365
	CER-MVS [66]	✓	0.359	0.305	0.332
	GeoMVSNet [140]	✓	0.331	**0.259**	**0.295**
	DUSt3R [111]	×	2.677	0.805	1.741
	Ours	×	3.261	1.170	2.266
	Ours	✓	2.930	1.265	2.098

Table 5. Quantitative evaluation of point clouds back-projected from multi-view metric depths on DTU. The methods are categorized into: a) traditional methods and b) learning-based MVS. The best results are in **bold**. Results of other methods are reported in DUSt3R [111].

Method	CLIP (Image)↑	Method	PSNR↑	SSIM↑	LPIPS↓
ImageDream [109]	83.77 ± 5.2	Zip-NeRF [3]	14.34	0.496	0.652
One2345++ [57]	83.78 ± 6.4	ZeroNVS [85]	17.13	0.581	0.566
IM3D [67]	**91.40 ± 5.5**	ReconFusion [115]	19.59	0.662	0.398
CAT3D [29]	88.54 ± 8.6	CAT3D [29]	**20.57**	**0.666**	**0.351**
Ours	88.76 ± 7.2	Ours	20.02	0.633	0.396

Table 6. Quantitative evaluation of monocular (left) and posed few-shot (right) 3d reconstruction. The best results are in **bold** and the second bests are underlined.

Figure 5. Monocular 3D reconstruction. Additional novel view renderings of our method are shown in the last two columns.

Input | ImageDream | CAT3D | Ours | View 1 | View 2

Here we mainly discuss DUSt3R and our method which regress 3D information from images by a network architecture without any 3D-specific operations. We found that the results of our method is better than DUSt3R for depth maps, but is worst for point cloud. Given that our method also achieves higher pose estimation accuracy, one possible reason is that DUSt3R is supervised directly by point positions. So it can achieve good point cloud evaluation, but fails for the two decoupled tasks. Overall, these two methods cannot achieve the same accuracy level as the methods with 3D domain knowledge embedded. But the results are accurate enough to serve as good 3DGS initialization.

4.4. 3D Reconstruction

Monocular. In the following, we evaluate the 3D reconstruction performance from single images. Specifically, we

Input | Reconfusion | CAT3D | Ours | GT | Ours

Figure 6. Sparse view 3D Gaussian Splatting reconstruction results from 3-view images input on CO3D dataset.

compare Matrix3D with diffusion-based optimization methods including ImageDream [109], One2345++ [57], IM-3D [67], and CAT3D [29] in terms of CLIP scores following CAT3D. Figure 5 and Table 6 illustrate the comparisons. Our method achieves comparable results to SOTA methods.

Sparse-view. We perform 3D reconstruction from sparse-view unposed images. Although previous methods have made similar attempts, they require given poses which is challenging to estimate for traditional SfM methods in the case of sparse view, and just use ground truth for exper-

| Unposed Sparse Input | Back-Projected Point Cloud | 3DGS Optimization | ◄── Novel View Renderings ──► |
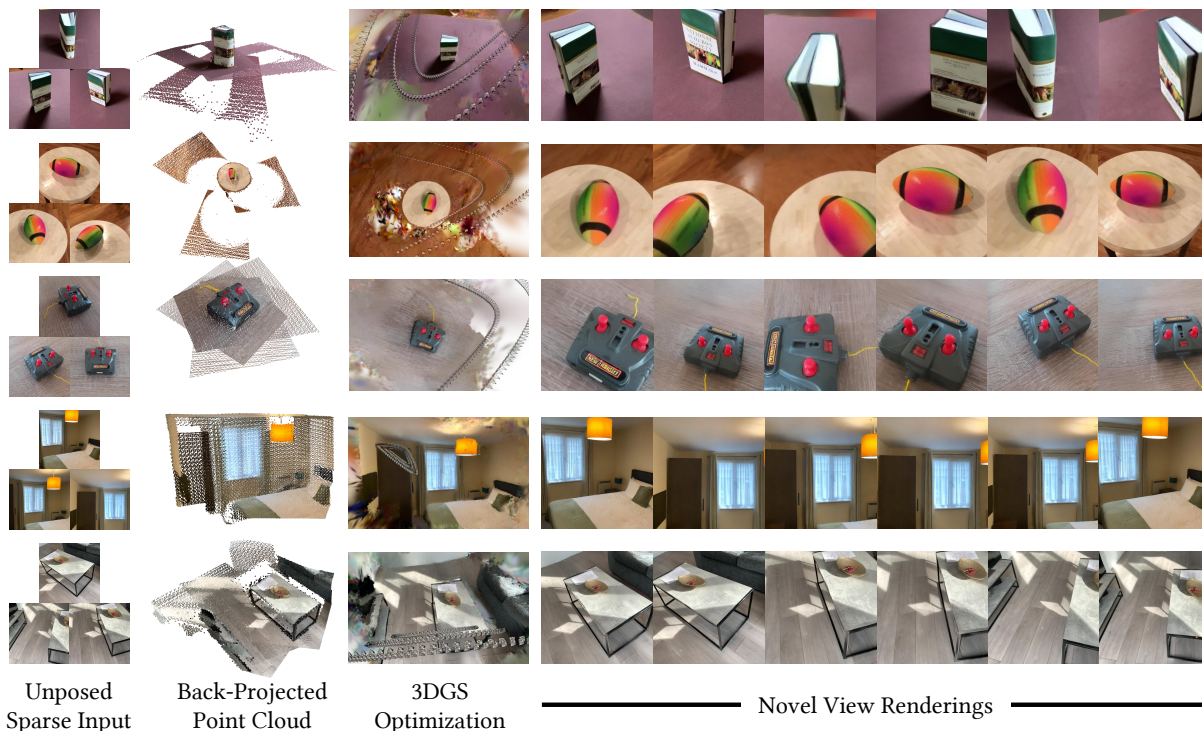
Figure 7. Unposed sparse-view 3D reconstruction results.

iments, leaving pose estimation problem unsettled. Our framework seamlessly integrates the pose estimation and reconstruction process into a single pipeline. Figure 7 shows the results of unposed 3-view images from CO3D and ARKitScene datasets. The reconstruction process can be found in Sec. 3.5. Results show that our method successfully performs the reconstruction given unposed images.

We also evaluate the task using ground truth poses as input. Following previous methods, we conduct 3-view reconstruction experiments on the same train/test split of the CO3D dataset, as done in CAT3D[29]. We use PSNR, SSIM, and LPIPS to evaluate performance (Table 6). Figure 6 shows the comparisons. Note that we use fewer than half of the novel views employed in CAT3D for reconstruction. Our method performs slightly worse than CAT3D, primarily due to the smaller number of views used for 3DGS training. Additionally, while our reconstruction method can be run on a single GTX 3090 GPU, CAT3D requires 16 A100 GPUs, making it impractical for most users.

4.5. Hybrid Tasks

One major advantage of multi-modal masked learning is that the model can accept flexible input combinations. If additional information is provided, the model knows how to take advantage of them, and thus produces better outputs. We show this quantitatively by adding depth ground truth to the NVS and the pose estimation tasks mentioned above. For pose estimation, as shown by Table 1, *Ours*

RGB+Depth consistently outperforms *Ours RGB Only* in terms of camera center accuracy. One possible reason is that depth information mitigates scale ambiguity. For NVS, as shown by Table 2, *Ours+Depth* also achieves better results than *Ours*. Intuitively, depth maps provide parital geometry, and thus facilitate NVS. In application, users can utilize depth measurement from active sensors to boost the performance of these two tasks.

5. Conclusion

In this paper, we introduced Matrix3D, a unified model that effectively addresses multiple photogrammetry tasks including pose estimation, depth prediction, and novel view synthesis using a multi-modal diffusion transformer (DiT). By employing a mask learning strategy, Matrix3D supports flexible input/output combinations, and maximizes training data from incomplete datasets. Through multi-round, multi-view, multi-model interactive generation, users can perform single or few-shot generation with one single model. Extensive experiments show that Matrix3D achieves SOTA performance in pose estimation and novel view synthesis tasks, showing its versatility on photogrammetry applications.

6. Acknowledgments

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. 6

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022. 3

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 7

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 4, 1, 2

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006. 2

[6] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Tat-Jun Chin, Chunhua Shen, and Ian Reid. Auto-rectify network for unsupervised indoor depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(12): 9802–9813, 2021. 3

[7] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3

[8] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*, pages 766–779. Springer, 2008. 2, 7

[9] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 3

[10] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023. 3

[11] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[12] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022. 2

[13] Kefan Chen, Noah Snavely, and Ameesh Makadia. Wide-baseline relative camera pose estimation with directional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3258–3268, 2021. 2

[14] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1538–1547, 2019. 2

[15] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 3

[16] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020. 7

[17] David J Crandall, Andrew Owens, Noah Snavely, and Daniel P Huttenlocher. Sfm with mrfs: Discrete-continuous optimization for large-scale structure from motion. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2841–2853, 2012. 2

[18] Hainan Cui, Xiang Gao, Shuhan Shen, and Zhanyi Hu. Hsfm: Hybrid structure-from-motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1212–1221, 2017. 2

[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. 3, 4, 1, 2

[20] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[21] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2

[22] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[23] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 6

[24] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8092–8101, 2019. 2

[25] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 2, 7

[26] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 2

[27] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2015. 2, 7

[28] Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *European Conference on Computer Vision*, pages 73–89. Springer, 2024. 3

[29] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 2, 3, 7, 8

[30] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, pages 11808–11826. PMLR, 2023. 3

[31] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 2, 7

[32] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[33] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19740–19750, 2023. 3

[34] Chris Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, pages 10–5244. Citeseer, 1988. 2

[35] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[36] 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. 3

[37] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. 2

[38] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[39] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[40] 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. *arXiv preprint arXiv:2311.04400*, 2023. 3

[41] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. 6

[42] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 3

[43] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. In *2024 International Conference on 3D Vision (3DV)*, pages 31–41. IEEE, 2024. 2

[44] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, page 2. Minneapolis, Minnesota, 2019. 3

[45] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[46] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38:199–218, 2000. 2

[47] Jeong-gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6775–6785, 2024. 3

[48] Maxime Lhuillier and Long Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):418–433, 2005. 2

[49] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 2, 3

[50] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer

for visual representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, 2021. 3

[51] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuandit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 4, 1, 2

[52] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023. 2, 6

[53] 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. 2

[54] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21136–21145, 2024. 3

[55] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5987–5997, 2021. 2

[56] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 2

[57] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10083, 2024. 3, 7

[58] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[59] 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. *arXiv preprint arXiv:2303.11328*, 2023. 3, 6

[60] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3, 6

[61] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 3, 6

[62] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[63] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[64] Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, and Yao Yao. Direct2.5: Diverse text-to-3d generation via multiview 2.5d diffusion. *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[65] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Aslfeat: Learning local features of accurate shape and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6589–6598, 2020. 2

[66] Zeyu Ma, Zachary Teed, and Jia Deng. Multiview stereo with cascaded epipolar raft. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 7

[67] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Natalia Neverova, Andrea Vedaldi, Oran Gafni, and Filippos Kokkinos. Im-3d: Iterative multiview diffusion and reconstruction for high-quality 3d generation. *arXiv preprint arXiv:2402.08682*, 2024. 3, 7

[68] David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[69] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 2

[70] 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. 1

[71] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3, 1

[72] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 4

[73] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[74] Alec Radford. Improving language understanding by generative pre-training. 2018. 3

[75] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[76] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the*

IEEE/CVF international conference on computer vision, pages 2349–2359, 2023. 3

[77] Rene 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. 3

[78] 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), 2022. 3

[79] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 4, 1, 2

[80] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 2

[81] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 4, 1, 2

[82] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[83] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 430–443. Springer, 2006. 2

[84] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 4

[85] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 3, 7

[86] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[87] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 6, 7

[88] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 7

[89] Philipp Schröppel, Jan Bechtold, Artemij Amiranashvili, and Thomas Brox. A benchmark and a baseline for robust multi-view depth estimation. In *2022 International Conference on 3D Vision (3DV)*, pages 637–645. IEEE, 2022. 7

[90] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International journal of computer vision*, 35:151–173, 1999. 2

[91] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[92] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 3

[93] Samarth Sinha, Jason Y Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B Lindell. Sparsepose: Sparse-view camera pose regression and refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21349–21359, 2023. 2

[94] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 2

[95] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024. 4

[96] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2

[97] 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. 5, 3

[98] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3

[99] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 3

[100] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint arXiv:2307.01097*, 2023. 3

[101] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018. 7

[102] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23:903–920, 2012. 2, 7

[103] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and

Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017. 7

[104] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024. 3

[105] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, 2021. 7

[106] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9065–9076, 2023. 3

[107] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 2, 6

[108] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21686–21697, 2024. 2

[109] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 7

[110] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 2, 3

[111] 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*, pages 20697–20709, 2024. 2, 3, 6, 7, 1

[112] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 3

[113] 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. 3

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

ings of the IEEE/CVF International Conference on Computer Vision, pages 17969–17980, 2023. 3

[115] 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. 2, 3, 7

[116] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv preprint arXiv:2405.14832*, 2024. 3

[117] Yuxi Xiao, Nan Xue, Tianfu Wu, and Gui-Song Xia. Levels2fm: Structure from motion on neural level set of implicit surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17205–17214, 2023. 2

[118] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4389–4398, 2024. 3

[119] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 6

[120] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *Proceedings of the AAAI conference on artificial intelligence*, 2020. 7

[121] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023. 3

[122] Jiayu Yang, Wei Mao, Jose M. Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[123] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7079–7088, 2024. 3

[124] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 6

[125] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 6

[126] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8574–8584, 2022. 7

[127] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2, 7, 1

[128] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5525–5534, 2019. 2

[129] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020. 6, 1

[130] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021. 3

[131] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 6

[132] 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. 4, 1, 2

[133] Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. In *European Conference on Computer Vision*, pages 163–179. Springer, 2024. 3

[134] Jingyang Zhang, Shiwei Li, Zixin Luo, Tian Fang, and Yao Yao. Vis-mvsnet: Visibility-aware multi-view stereo network. *International Journal of Computer Vision*, 131(1): 199–214, 2023. 2, 7

[135] Jingyang Zhang, Shiwei Li, Yuanxun Lu, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, and Yao Yao. Jointnet: Extending text-to-image diffusion for dense distribution modeling. *International Conference on Learning Representations (ICLR)*, 2024. 3

[136] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, pages 592–611. Springer, 2022. 2

[137] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024. 2, 4, 6, 1

[138] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. *arXiv preprint arXiv:2404.19702*, 2024. 3

[139] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3

[140] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. Geomvsnet: Learning multi-view stereo with geometry perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21508–21518, 2023. 7

[141] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 4, 1, 2