Multi-view Reconstruction via SfM-guided Monocular Depth Estimation

Haoyu Guo[1*] He Zhu[2*] Sida Peng[1] Haotong Lin[1] Yunzhi Yan[1]
Tao Xie[1] Wenguan Wang[1] Xiaowei Zhou[1] Hujun Bao[1†]
[1]Zhejiang University [2]Beijing Normal Univeristy
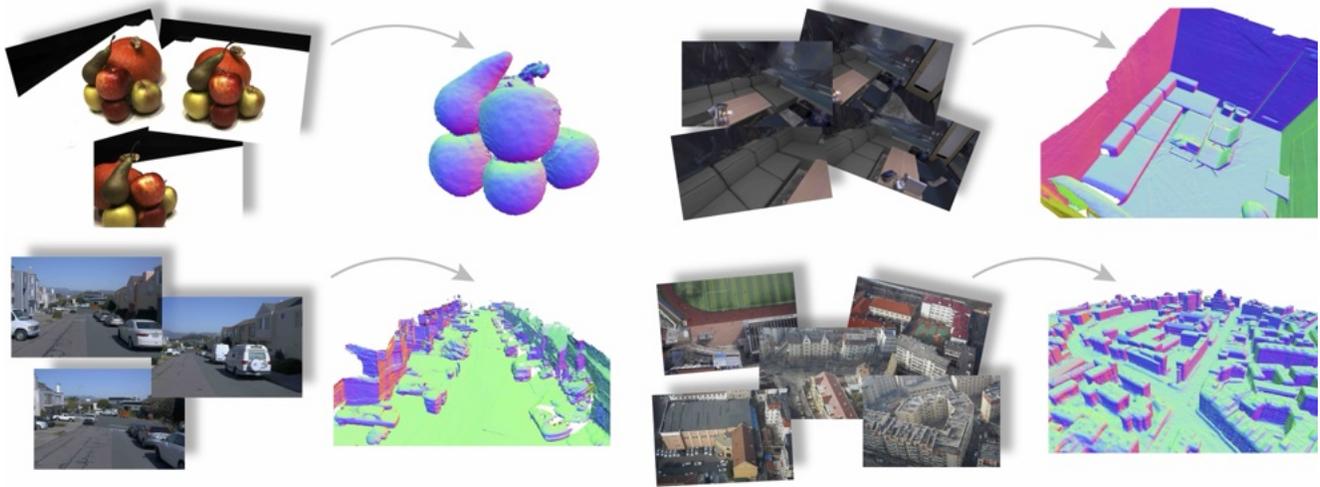
Figure 1. We propose Murre, a new method for multi-view 3D reconstruction based on SfM-guided monocular depth estimation. Building on Stable Diffusion [48], Murre demonstrates exceptional generalization capabilities after fine-tuning on a modest amount of synthetic data. Murre is capable of achieving high-quality reconstructions for a variety of real-world scenarios, including object-level, indoor, street, and aerial scenes.

Abstract

This paper aims to reconstruct the scene geometry from multi-view images with strong robustness and high quality. Previous learning-based methods incorporate neural networks into the multi-view stereo matching and have shown impressive reconstruction results. However, due to the reliance on matching across input images, they typically suffer from high GPU memory consumption and tend to fail in sparse view scenarios. To overcome this problem, we develop a new pipeline, named Murre, for multi-view geometry reconstruction of 3D scenes based on SfM-guided monocular depth estimation. For input images, Murre first recovers the SfM point cloud that captures the global scene structure, and then use it to guide a conditional diffusion model to produce multi-view metric depth maps for the final TSDF fusion. By predicting the depth map from a single image, Murre bypasses the multi-view matching step and naturally resolves the issues of previous MVS-based methods. In addition, the diffusion-based model can easily leverage the powerful priors of 2D foundation models, achieving good generalization ability across diverse real-world scenes. To obtain multi-view consistent depth maps, our key design is providing effective guidance on the diffusion model through the SfM point cloud, which is a condensed form of multi-view information, highlighting the scene's salient structure, and can be readily transformed into sparse depth maps to drive the image-space estimation process. We evaluate the reconstruction quality of Murre in various types of real-world datasets including indoor, streetscapes, and aerial scenes, surpassing state-of-the-art MVS-based and implicit neural reconstruction-based methods. The code and supplementary materials are available at https://zju3dv.github.io/murre/.

---

* Equal contribution
† Corresponding authors

1. Introduction

Multi-view 3D reconstruction aims to recover the accurate scene geometry from multiple images with known camera poses, which has extensive applications in robotics, autonomous driving, and virtual reality. Traditional multi-view stereo algorithms [51, 53, 54, 56] and neural reconstruction methods [63, 75, 77] define an optimization target based on multi-view photometric consistency to reconstruct 3D scenes from the image signal. Though these methods achieve impressive performance, they struggle with reconstruction in low-textured regions, as there might be several plausible solutions satisfying the photometric constraint in such area. Moreover, the reconstruction speed of these methods is typically slow due to the time-consuming optimization process.

With the development of deep learning techniques, some recent methods [59] introduced neural networks into the multi-view stereo reconstruction framework for incorporating the 3D scene prior that is learned from the large-scale data. These methods generally extract informative features from images, aggregate multi-view features in 3D space, and determine whether a point is on the 3D surface by implicitly considering the multi-view feature consistency. With the learned scene prior, these methods resolve the ambiguity in low-textured regions and enable the feedforward inference with a faster reconstruction speed.

However, such learning-based MVS methods [4, 18, 64, 67, 72–74] still face three challenges. First, due to the high spatial complexity of 3D space, aggregating features in 3D space tends to consume a large amount of GPU memory, thereby limiting the resolution of the reconstructed geometry. Second, these methods implicitly rely on the inductive bias of multi-view consistency, thus tending to fail in a sparse view setting where there are many regions that cannot be matched across input images. Third, these methods generally require ground-truth 3D data for training. Because high-quality 3D data is limited, it is difficult to train them on a large amount of data, resulting in a limited generalization ability.

In this paper, we propose a novel learning-based framework, named Murre, that combines SfM techniques with monocular depth estimators to reconstruct 3D geometry from multi-view images. Our core innovation lies in injecting SfM priors into diffusion-based depth estimation to produce multi-view consistent depth maps for input images, achieving 3D reconstruction with strong generalization capability and high quality. Specifically, given input images, Murre first recovers a sparse point cloud through an off-the-shelf SfM method [22, 52], which captures the scene structure globally. Then, the point cloud acts as a guidance for a conditional diffusion model, which takes the projected sparse depth map and an RGB image as input and predicts the corresponding metric depth map. Finally, the 3D geom-etry of the target scene is obtained by performing the point cloud fusion [17, 40] or TSDF fusion [42] on metric depth maps of input images. Murre has two advantages over previous learning-based MVS methods [18, 67, 73, 74]. First, by utilizing an SfM-guided diffusion model to predict the depth map from a single image, Murre bypasses the multi-view matching step, thereby addressing the issues of high memory consumption and difficulty in handling sparse view scenarios. Second, Murre adapts the diffusion model from a pre-trained stable diffusion model, making a clever use of the powerful priors of the 2D foundation model. Our experimental results in Section 5 show that fine-tuning on a small amount of synthetic 3D data can achieve good generalization results in diverse real-world scenes, including object-level, indoor, street and aerial scenes.

More specifically, a key design of Murre is exploring the explicit intermediate representation, SfM point clouds, to guide the conditional diffusion model to produce multi-view consistent depth maps with an accurate scale. As a condensed form of multi-view information, SfM point clouds effectively highlight the scene's salient structure and can be readily transformed into sparse depth maps, which serve as conditional signal well-suited for the image-space estimation process. When predicting the depth map of an input image, we first project the reconstruction SfM point cloud onto the corresponding view. To perform the dense guidance on the diffusion model, we densify the sparse depth map using the k-nearest neighbors algorithm and compute a corresponding distance map as a confidence indicator. The diffusion model takes the RGB image, densified depth map, and distance map as input, and predicts the target metric depth map. Experimental results show that our strategy significantly enhances the multi-view consistency of the predicted depth maps from the diffusion model.

We evaluate Murre on several reconstruction benchmarks of diverse types, including DTU [27], ScanNet [9], Replica [57], Waymo [60] and UrbanScene3D [33] datasets. Murre consistently outperforms state-of-the-art monocular depth estimation [29, 70, 71] methods and achieves competitive performance with recent Multi-view Stereo and implicit neural reconstruction-based method [50, 59, 77]. Detailed ablations and analysis on the SfM method and depth conditioning designs are included to demonstrate the robustness and efficiency of our design.

In summary, our contributions are as follows:

- We propose a novel approach to inject SfM priors into diffusion-based depth estimation, enabling highly accurate and multi-view consistent depth predictions for each viewpoint.
- Based on the proposed depth estimator, we design a new multi-view 3D geometry reconstruction framework and process some synthetic datasets to facilitate training.
- We evaluate our method on diverse real-world scene data,

including objects, indoor, streetscapes, and aerial scenes, demonstrating the superior performance and generalization capability of our approach.

2. Related works

2.1. Multi-view reconstruction

Multi-view 3D reconstruction has been a long focus in the field of 3D computer vision for decades. Traditional multi-view stereo (MVS) methods [16, 26, 51, 53, 54, 56] primarily rely on photometric consistency to perform dense matching between images from different views to estimate depth, leading to a subsequent reconstruction of the scene. However, limited by their weak feature representation capability, they struggle to deliver high-quality reconstruction results in challenging scenes.

Recently, with the rise of deep learning, learning-based MVS methods have been widely explored. These methods typically construct a 3D cost volume and utilize 3D CNNs [18, 67, 73, 74] or transformers [4, 64, 72] to predict per-view depth map, and then fuse all views together. However, the introduction of 3D cost volume results in massive computational overhead and still struggles to provide reasonable reconstruction results under sparse viewpoints or textureless situations. In contrast, NeuralRecon [59] directly constructs a feature volume in the world coordinate system and predicts a TSDF volume for scene reconstruction. Although these approaches can handle low-textured regions, they require expensive 3D ground truth as supervision and lacks reconstruction details. Optimization-based neural reconstruction approaches [31, 63, 75, 77] leverage RGB supervision to implicitly optimize the 3D scene representation, which alleviates the requirement for accurate 3D ground truth. However, they still necessitate time-consuming optimization for each distinct scene. Our method proposes a new multi-view reconstruction framework which combines monocular depth estimation with sparse SfM point clouds, addressing the challenges posed by learning-based methods in handling sparse viewpoints and weak textures.

2.2. Monocular depth estimation

Monocular depth estimation, which aims to directly predict a dense depth map from one single image, holds expansive application value in areas such as generative AI [24, 48] and autonomous driving [14, 35], specifically, they are shown to provide a strong prior for multi-view reconstruction [7, 61, 65, 77]. However, recovering metric scale depth purely from one single image is an ill-posed problem. Previous metric monocular depth estimation methods [1–3, 78] learn the metric scale by training on specific scene data, but they essentially overfit on the training scenarios and face difficulties in generalizing to very different scenes. Recent works [12, 76] boost the generalizability by explic-

itly modeling different camera models through encoding or constructing a canonical camera model. Although these methods improve the performance on depth estimation, they still struggle to produce multi-view consistent depth maps for observed images of a target 3D scene.

Alternatively, affine-invariant relative monocular depth estimation methods predict the relative relationships between pixels without a specific metric scale. [46, 70, 71] achieve better generalization capabilities through pre-training on large-scale datasets. Recently, diffusion-based methods [15, 19, 29] repurpose the diffusion [23] framework to denoise the depth map from the image being used as the condition. With the help of the powerful priors provided by Stable Diffusion [48], they fine-tune models with decent generalization on a minor quantity of synthetic datasets. However, these methods lack global metric information and struggle to guarantee the internal relative scale consistency of a single depth map, making them less applicable for reconstruction. Additionally, [25, 30, 38, 55, 69] explore estimating consistent depth from monocular video.

2.3. Depth completion

Depth completion [8, 32, 37, 66, 79] aims to recover dense pixel-wise depth maps from sparse depth maps obtained via SfM or sensors. Traditional methods have made extensive attempts to incorporate multi-modal data fusion into the completion pipeline, such as RGB images [28, 41, 45, 81], semantic maps [39], and normal maps [68]. Building upon it, later methods focus on refining the network architectures, such as different convolution structures [5, 6, 34, 36], and training strategies [79]. However, they neglect the inner properties of the non-uniform sparse depth points. [8, 37, 66] try to address this by iteratively merging or locality-aware spatial propagation. Although they achieve quality improvement, these methods are all premised on the assumption of accurate sparse depth, and struggle to yield satisfactory results when tasked with noise-containing depths such as SfM depth.

3. Method

Given multi-view images of the target scene along with sparse reconstruction results from Structure from Motion (SfM), our goal is to utilize the SfM priors to estimate the dense depth for each view, which can be fused to obtain the geometry reconstruction of the scene, as illustrated in Figure 2. We build our depth estimator based on a conditional diffusion model (Section 3.1). To leverage multi-view information, we propose an explicit intermediate representation to incorporate SfM priors during the depth estimation process (Section 3.2). Based on our proposed depth estimation method, we design a pipeline to reconstruct the scene geometry from input multi-view images (Section 3.3). Finally, we elaborate on our training details in Section 3.4.
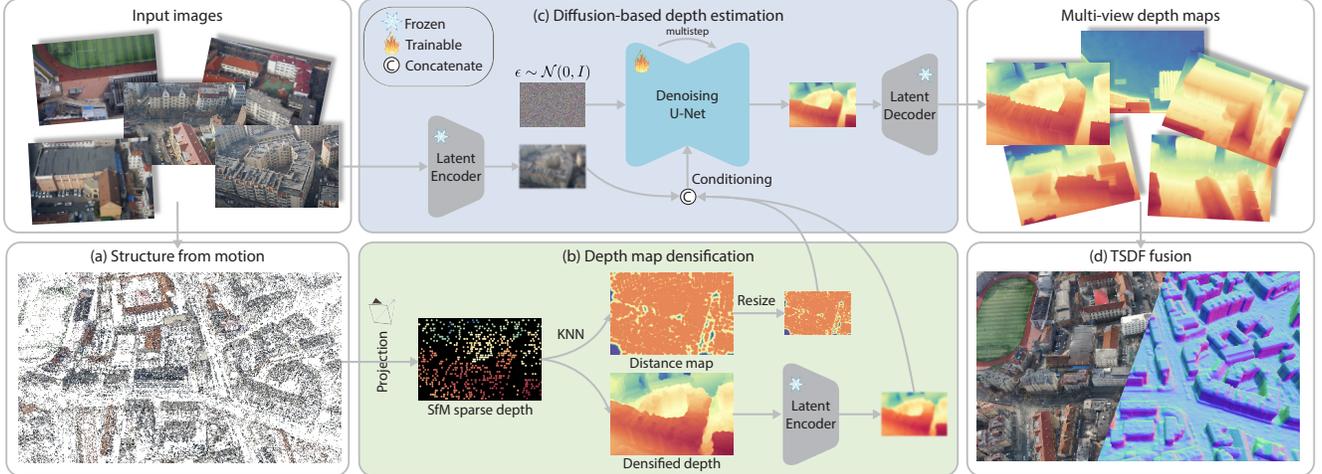
Figure 2. **Overview of our multi-view reconstruction pipeline.** Given multi-view images, we first employ a Structure from Motion (SfM) method [22] to derive sparse 3D scene structures (a). These 3D structures are then encoded into an intermediate explicit representation (b), which is used as a condition for depth estimation (c). Finally, we conduct a TSDF fusion [42] to achieve the final reconstruction (d).

3.1. Diffusion-based depth estimation

Recent multi-view stereo methods [50, 73, 80] typically rely on multi-view image matching to build cost volumes, which makes them difficult to handle textureless regions and sparse viewpoints. To address this, we propose injecting multi-view information into monocular depth estimation [29] to avoid the need for multi-view image matching. To achieve this goal, we condition our method on a diffusion-based monocular depth estimation model, which has shown strong performance benefiting from powerful modern image diffusion models. In the following, we first introduce the prior knowledge of the image diffusion model and then explain how to use it to facilitate depth estimation.

Image generation field has recently made significant progress with large-scale data and diffusion models. A representative method is Denoising Diffusion Probabilistic Models (DDPM) [23], which learns the reverse process of adding noise and generating a clean image from a random noise. To achieve high-resolution image generation, [48] suggests conducting the noise addition and denoising process in the latent space. A Variational Autoencoder (VAE) [11] is first trained on a very large dataset of images, enabling bidirectional transformations between image space and latent space. [49] introduces a conditioning mechanism to diffusion models, enabling control over the generated outcomes based on textual or visual information.

Inspired by [29], we formulate depth estimation as a latent conditional diffusion process. Specifically, in the forward process, the Gaussian noise is gradually added to the depth d to obtain the noisy depth sample d_t, starting from $d_0 = d$ to levels $t \in \{1, \ldots, T\}$ as the following equation:

$$d_t = \sqrt{\bar{\alpha}_t} d_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \qquad (1)$$

where $\epsilon \sim \mathcal{N}(0, I)$, $\bar{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$, and $\{\beta_1, \ldots, \beta_T\}$ is the variance schedule of a process with T steps. In the reverse process, a learned denoiser is used to predict the noise ϵ from the monocular image condition, which is fed into the denoising equation [23] to calculate the depth map d_{t-1} from d_t. Finally, we iteratively perform the denoising process and obtain the clean latent d_0, which is converted into the depth through the decoder of the VAE.

3.2. Injecting SfM priors into depth estimator

Previous methods [29] achieve high-quality monocular depth estimation by using the diffusion method with monocular images as a condition. However, directly fusing their output depth produces poor reconstruction results because the model struggles to predict multi-view consistent depth maps with only monocular images due to the ill-posedness. To achieve high quality multi-view 3D reconstruction, an intuitive solution is to incorporate multi-view information into the depth estimation model. However, directly incorporating information from multi views under the diffusion framework is quite challenging, because the number of input views and the distribution of their camera poses are unknown, resulting in a large variance in input data. Moreover, the image-space estimation process makes it difficult to fuse multi-view image features.

To overcome these problems, we propose to use the SfM point cloud as an explicit intermediate representation, enabling effective and efficient injection of multi-view information into the monocular depth estimation model. Using the SfM point cloud as a condition has several advantages: Firstly, SfM points can be regarded as a condensed form of multi-view information and scene 3d structure, which can guide the prediction of the depth maps. Secondly, the research on SfM has been conducted for nearly two decades,

making it highly mature and robust [22, 43, 52, 58]. Finally, SfM point clouds can be projected into each view according to the visibility as sparse depth maps, making them well-suited for the image-space estimation task.

Since the projected depth maps are sparse and not suitable for image-space conditioning, we propose to first densify them through the k-nearest neighbors (KNN) interpolation. Specifically, for each pixel p in a sparse depth map that lacks a value, we find the k nearest pixels with values based on Euclidean distance. We then perform a weighted average of these k pixel depths using the inverse of the distances as weights to assign a value to pixel p. Using this method, we densify the depth map. After that, we pass the densified depth map through an encoder to serve as a condition. To help the network to distinguish between original and interpolated pixels and better utilize the SfM information, we use an additional map that records the Euclidean distance from each pixel to its nearest neighbor with a value. The distance map is used as an additional condition.

Due to the significant variations in the range of SfM-projected depth values across different views of various scenes, we first normalize the spase SfM depth maps. Considering that SfM points often contain outliers, we initially filter out the top and bottom 2% of depth values for each view. Given that SfM projections only cover a small portion of the pixels and the actual depth range of the entire image maybe greater than the depth range of the SfM points, we expand the calculated range based on SfM depth. Specifically, we adjust the minimum to 0.8 times its original value and the maximum to 1.2 times. This adjustment ensures that the entire image's depth range is well covered in most cases, thereby minimizing the impact of edge cases on training. For the ground truth depth map, we use the same range for normalization and then map it to the latent space to supervise the denoising process. This design ensures that the trained model can predict depths on the same scale as the input SfM depth, which can then be effectively applied to subsequent multi-view 3D reconstruction.

3.3. Multi-view 3D reconstruction

Given input multi-view images, we first reconstruct sparse point clouds using SfM techniques. Based on the SfM point cloud, we predict depth maps for the input images. Specifically, we map the image and SfM information used as conditions to the latent space following the method described in Section 3.2 to obtain the normalized depth map. We use the recorded normalization range of the SfM depth to denormalize the predicted depth back to the original scale of the SfM depth. After that, we use off-the-shelf multi-view depth fusion methods such as point cloud fusion [17, 40] or TSDF fusion [42] to obtain the scene's 3D geometry.

We found that directly fusing the predicted depth maps yields suboptimal results. One reason is a slight incon-

sistency in the scale between the predicted depth and the SfM depth. To overcome this issue, we further align the predicted depth map with the SfM sparse depth using RANSAC based linear regressor. Specifically, RANSAC iteratively identifies inliers and estimates the optimal scale and shift parameters by minimizing the reprojection error between the predicted depth and the SfM depth. In each iteration, a subset of points is randomly sampled to compute a candidate transformation, which is then evaluated based on the number of inliers it produces. The final transformation is selected as the one with the highest consensus, ensuring robust alignment by effectively handling outliers and adjusting for both translation and scaling discrepancies.

3.4. Training

To train our model, we first run the SfM algorithm to reconstruct sparse point clouds of training scenes from input multi-view images. Specifically, we choose Detector-free SfM [22], which is based on a detector-free matcher [10, 58]. This method is better suited for handling areas with weak textures compared to classical SfM approaches [52]. After running SfM on each scene, we project the obtained sparse point cloud according to the visibility to get the sparse depth map for each view.

Based on the constructed training data, we train a depth estimator that uses both images and SfM depth as conditions. To enable robust generalization capabilities, instead of training from scratch, we use the pretrained Stable Diffusion V2 model [48] as the initialization, and we fix the VAE and only finetune the UNet during the training. We normalize the GT depth map of each frame according to Section 3.2, then replicate it across three channels to use the VAE's encoder to map it into the latent space. We then add noise to the latent code of the GT depth and train the denoiser based on this. In addition to the latent codes of the image and densified SfM depth map, we also utilize the Euclidean distance map as condition. Because the information of distance map is of low frequency, instead of passing it through the encoder, we directly resize it to the same resolution as the latent code. We concatenate them together with the noise and send them into the UNet to predict $\hat{\epsilon}$. The training loss is an mse loss defined as:

$$\mathcal{L} = \|\epsilon - \hat{\epsilon}\|^2,\tag{2}$$

where ϵ is the noise added to the latent code of GT depth.

4. Implementation details

We employ the official pre-trained model of Stable Diffusion v2 model [48] as our foundation, with the Variational Autoencoder (VAE) fixed and only the U-Net fine-tuned. Our framework is implemented using PyTorch Lightning [13] and diffusers [62]. During the training stage, we utilize a mixture of data from Hypersim and 3D Ken Burns,

and for alignment purposes, we downscale the resolution of Hypersim from 768×1024 to 384×512, and center-cropped the 3D Ken Burns data from 512×512 to 384×512. We train our model using eight Nvidia Tesla A100 GPUs with a batch size of 128 across 50 epochs (about 25 hours), employing the Adam optimizer with a learning rate of 3×10^{-5}. To enhance the robustness of depth estimation, inspired by [29], we employ a test-time ensemble approach, performing inference 5 times and then taking the pixel-wise median.

5. Experiments

5.1. Datasets and metrics

5.1.1 Training datasets

We trained our model on two synthetic datasets. (1) **Hypersim** [47] is a photorealistic synthetic dataset consisting of 461 indoor scenes. Based on the official split, we filter out broken or low-quality scenes from the 365 training scenes, retaining 345 scenes with 56,077 frames. For each scene, we performed triangulation based on the provided camera parameters to obtain sparse reconstruction results. (2) **3D Ken Burns** [44] is a large-scale synthetic dataset, containing 23 in-the-wild scenes with two sequences each. Due to the unknown camera parameters in this dataset, we employ Detector-free SfM [22] for calibration and sparse reconstruction. After excluding sequences of poor quality or those where SfM does not work properly, we retained 12 sequences with 30,309 frames for training.

5.1.2 Evaluation datasets.

To validate the generalizability and robustness of our multi-view reconstruction method across different data types, we conduct evaluations on various types of real-world data. (1) Object-level scenes: **DTU** [27] DTU comprises 80 object-level scenes captured using multiple cameras under various lighting conditions. We use 15 scenes for 3-view reconstruction evaluation, following [77]. (2) Indoor scenes: **ScanNet** [9] is a RGBD dataset containing 1,613 indoor scenes. We evaluate on 4 scenes in validation set, following [20, 77]. **Replica** [57] includes 18 indoor scenes with high quality Ground Truth geometry at room scale. We evaluate on 3 test scenes according to the split in [77]. (3) Streetscapes: **Waymo Open Dataset** [60] is an autonomous driving dataset collected with five LiDAR sensors and five high-resolution pinhole cameras. We select 32 static sequences for evaluation, following [21]. (4) Aerial views: **UrbanScene3D** [33] is an outdoor dataset captured using drones, each scene with 2500-6000 images. We evaluate on two scenes covering 10^6 square meters. For DTU, ScanNet, Replica, and Waymo, we conduct sparse reconstructions using the triangulation mode of Detector-free SfM from known camera parameters. For UrbanScene3D, we

Mairgold **Depth-Anything** **IGEV-MVS** **DUSt3R** **Ours**

Figure 3. **Qualitative comparison on DTU dataset.** Except for DUSt3R which directly infers the point cloud, all other methods use Fusibile [17] to fuse depth maps, filtering out points with inconsistencies across different viewpoints. Murre produces a more complete final point cloud due to its relatively accurate multi-view consistent depth estimation. Please refer to the supplementary materials for results on other datasets.

obtain the calibration as well as sparse reconstruction results by COLMAP.

5.1.3 Metrics

Given our focus on multi-view geometric reconstruction, we primarily evaluate reconstruction quality. We report Chamfer distance on DTU following [77], F-score with a threshold of 5cm on ScanNet and Replica following [77]. As for Waymo, we report the RMSE (root mean square error) between predicted depth maps and the provided lidar depth in the image space to represent geometry quality due to the lack of 3D Ground Truth geometry, following [21].

5.2. Comparison

Murre is compared with recent methods on monocular depth estimation, depth completion, and multi-view reconstruction, which are introduced in the supplementary material. We report both qualitative and quantitative comparisons against baselines, showcasing the reconstructed geometry in Figure 3 and evaluating the reconstruction results on DTU in Table 1. Please refer to more quantitative results on street and aerial scenes in the supplementary material, as well as additional qualitative results (including predicted depth maps and reconstructed geometry). It is evident that Murre consistently outperforms the baseline approaches across various types of real-world datasets. While Marigold and Depth-Anything produce visually appealing depth visualizations, the depth values are often inaccurate on many datasets, leading to poor reconstruction results. Metric3D, trained on a vast amount of real-world streetscape data, performs well on the Waymo dataset but shows poor performance on other datasets. Despite align-

Table 1. **Quantitative comparison on DTU dataset**. We evaluate Chamfer Distance of each method on each scene (measured in millimeters) and highlight the top three methods in red, orange, and yellow, respectively. Apart from metrics, we also report the amount of training data used by each method. Please refer to the supplementary materials for results on other datasets.

	COLMAP	RealityCapture	VolSDF	MonoSDF	Marigold	Depth-Anything	Depth-Anythingv2	Metric3D	MVSNet	IGEV-MVS	DUSt3R	Ours
scan24	4.45	4.19	5.24	3.47	4.77	2.99	4.07	5.99	3.16	4.20	2.24	1.31
scan37	4.67	3.85	5.09	3.61	7.61	3.77	4.33	5.26	4.36	5.05	3.43	3.14
scan40	2.51	2.26	3.99	2.10	5.34	3.01	2.54	4.13	2.59	3.47	2.77	1.53
scan55	1.90	2.49	1.42	1.05	4.75	2.10	4.07	4.22	2.35	2.79	2.40	0.92
scan63	2.81	3.49	5.10	2.37	4.07	2.98	3.05	3.14	1.88	3.05	3.71	1.19
scan65	2.92	3.97	4.33	1.38	4.15	4.10	3.61	5.00	3.28	3.16	3.22	1.98
scan69	2.12	1.91	5.36	1.41	7.38	2.73	3.16	4.20	2.32	3.16	2.49	1.25
scan83	2.05	2.49	3.15	1.85	6.00	2.38	3.29	4.78	1.88	3.17	3.34	2.16
scan97	2.93	2.37	5.78	1.74	5.38	4.17	4.86	4.54	1.87	3.62	2.71	1.40
scan105	2.05	2.27	2.07	1.10	5.26	3.45	2.79	4.85	2.06	4.16	2.84	1.07
scan106	2.01	2.90	2.79	1.46	5.36	4.40	5.24	5.44	2.25	3.09	2.69	1.26
scan110	N/A	4.60	5.73	2.28	5.51	1.97	5.29	4.69	1.92	5.97	3.08	0.95
scan114	1.10	1.38	1.20	1.25	4.65	2.53	2.41	7.82	1.40	2.49	2.20	0.82
scan118	2.72	2.57	5.64	1.44	5.63	2.94	3.01	4.36	2.43	3.23	2.52	1.18
scan122	1.64	1.76	6.20	1.45	5.99	2.88	3.57	6.78	1.94	1.90	2.51	1.07
Average	2.56	2.84	4.21	1.86	5.46	3.09	3.69	5.01	2.38	3.50	2.81	1.42
Training Data	-	-	-	-	74K	1.5M	0.6M	8M	27.1K	35.8K	8.5M	86.4K

ing with SfM depth on these datasets, the results remain low quality. MVSNet and IGEV-MVS, although trained on DTU, exhibit only moderate performance when tested on DTU with inputs of three views. NeuralRecon and SimpleRecon, both trained on ScanNet, demonstrate good results when evaluated on ScanNet, but we find that they nearly fail to work on other datasets.

5.3. Ablation studies

5.3.1 Trade-off between precision and speed

Murre incorporates several factors that impact speed and reconstruction quality: 1) the number of denoising steps, 2) the number of ensemble size, 3) the implementation of additional alignment. To analyze the influence of these factors, we conduct a series of experiments on Replica dataset, the results are shown in Table 2. For fairness, all tests are conducted on a single A6000 GPU with the batch size as 1. The actual speed of Murre can be further improved by increasing the batch size. For the experiment where $N_{step} = 1$, we first distill our model using latent consistency model (LCM) (details in supplementary materials) to enable it to achieve satisfactory results with just one step.

The results indicate that: 1) with the LCM-distilled model, effective depth estimation can be achieved with just one denoising step, significantly reducing denoising time without sacrificing much reconstruction quality.; 2) reducing the ensemble size also considerably enhances speed with minimal compromise in reconstruction quality; 3) The time consumed by alignment is minimal, while we also

Table 2. **Trade-off between precision and speed.** We evaluate the inference speed (in seconds) of each view (at the resolution of 680x1200) and final reconstruction quality of Murre under different settings. This allows us to select the most suitable setting based on specific application requirements.

N_{Step}	$N_{Ensemble}$	Alignment	Inference time $\downarrow$	F-Score $\uparrow$
10	5	RANSAC	12.166	0.853
10	1	RANSAC	2.969	0.850
1	5	RANSAC	3.202	0.829
1	1	RANSAC	0.840	0.828
1	1	Least Square	0.836	0.812
1	1	No Alignment	0.829	0.780

demonstrate that since Murre can predict depth maps at the same scale as the conditioned sparse depth maps, the alignment step is not indispensable. In summary, Murre offers a flexible trade-off between speed and reconstruction quality, allowing for adjustments based on specific application requirements.

5.3.2 SfM method

Our training data are all processed by DF-SfM [22] for sparse reconstruction, primarily due to DF-SfM's robustness in scenes with low textures. During inference, other SfM methods can be utilized. Note that our experiments on Urbanscene3D are based on COLMAP rather than DF-SfM. To further analyze the impact of the choice of SfM

Table 3. **Quantitative ablations of SfM methods.** While Murre is trained with the sparse points from DF-SfM (LoFTR), it is capable of directly utilizing the condition from COLMAP and DF-SfM (DKM) for inference, demonstrating the robustness.

	COLMAP	DF-SfM (LoFTR)	DF-SfM (DKM)
F-score	0.645	**0.853**	0.842
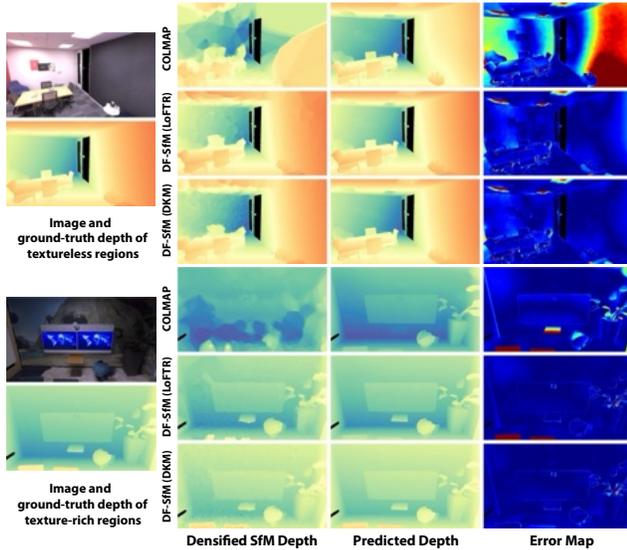
Figure 4. **Qualitative ablations of SfM methods.** We conduct ablations on both texture-rich and textureless scenes.

method on our performance, we conduct an ablation study on Replica, which is an indoor dataset and contain large areas of low texture in some scenes, the results are shown in Table 3 and Figure 4. During the construction of training data, DF-SfM uses LoFTR as its matcher, and for inference, good results can be achieved whether we continue using LoFTR or switch to DKM. Meanwhile, the performance of Murre based on COLMAP depends on specific characteristics of the scene. In regions with weak textures, COLMAP's keypoints tend to be extremely sparse and noisy, as a result, while the predicted depth map may look visually appealing, it often contains substantial errors. However, in regions with relatively rich textures, although COLMAP's keypoints are sparser than DF-SfM, they remain within an acceptable range and have good accuracy, resulting in high-quality depth maps generated by Murre. In overall, Murre is robust to the choice of SfM method: as long as the SfM can produce a reasonably good sparse point cloud for the scene, we can achieve high-quality depth estimation and reconstruction results based on it.

5.3.3 Depth conditioning

To introduce SfM priors into depth estimation, we first densify the SfM-projected depth using KNN, then pass it through an encoder and concatenate it with a downsampled

Table 4. **Ablation studies of depth conditioning.** We evaluate the impact on performance when using different values of k in KNN (where $k = 0$ indicates no densification) and the effect of using the distance map as condition.

k	0	0	1	1	3	3
distance map	✘	✔	✘	✔	✘	✔
F-score	0.543	0.528	0.783	0.840	0.753	**0.853**

distance map as condition. To analyze the impact of this design, we conduct ablation studies on KNN and the distance map on Replica dataset. As shown in the Table 4, the performance is poor without KNN, suggesting that original sparse depth map is not suitable for directly inputting into the encoder. A slight improvement is observed with $k = 3$ compared to $k = 1$, indicating that $k = 3$ provides more information. Furthermore, whether at $k = 1$ or $k = 3$, omitting the distance map as a condition significantly worsens the results, demonstrating that the distance map effectively conveys crucial information about which pixels are originally valued and which are interpolated.

6. Conclusion

We propose a new paradigm for multi-view geometric reconstruction of 3D scenes based on monocular depth estimation. Traditional routines for image-based 3D reconstruction rely on either multi-view stereo matching or differentiable inverse rendering which may suffer from high computational cost and low reconstruction quality caused by insufficient textures or very sparse input views. Recent advances in large vision models have shown remarkable capabilities to recover high-quality depth maps from single images, but the estimated depth values are often imprecise and inconsistent across different views due to the inherent single-view ambiguity. We propose to incorporate SfM point clouds as a condition into the diffusion-based depth estimation process, thus naturally enhancing the metric accuracy and multi-view consistency of the estimated depth maps, which are then directly fused to reconstruct the scene geometry. We evaluate the reconstruction quality in various types of real-world datasets including indoor scenes, streetscapes, and aerial views, surpassing state-of-the-art MVS and implicit neural reconstruction-based methods.

Murre also has some known limitations. First, we require running SfM, which is infeasible in extreme cases where only two views with minimal overlap are available, making it nearly impossible to estimate camera poses or obtain a sparse point cloud, thereby preventing further reconstruction. Additionally, Murre is designed to reconstruct static scenes and is hard to handle moving elements. Future work could explore integrating DUSt3R-like approaches to address more extreme cases and incorporating tracking methods to handle dynamic scenes.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021. 3

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *ECCV*, 2022.

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3

[4] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer: Multi-view stereo by learning robust image features and temperature-based depth. *TMLR*, 2023. 2, 3

[5] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *ECCV*, 2018. 3

[6] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *AAAI*, 2020. 3

[7] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *CVPR*, 2024. 3

[8] Andrea Conti, Matteo Poggi, and Stefano Mattoccia. Sparsity agnostic depth completion. In *WACV*, 2023. 3

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2, 6

[10] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *CVPR*, 2023. 5

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 4

[12] Jose M Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. Camconvs: Camera-aware multi-scale convolutions for single-view depth. In *CVPR*, 2019. 3

[13] W.A. et al. Falcon. Pytorch lightning. https://github.com/PytorchLightning/pytorch-lightning, 2019. 5

[14] Michaël Fonder, Damien Ernst, and Marc Van Droogenbroeck. M4depth: Monocular depth estimation for autonomous vehicles in unseen environments. *arXiv preprint arXiv:2105.09847*, 2021. 3

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

[16] Simon Fuhrmann, Fabian Langguth, and Michael Goesele. Mve-a multi-view reconstruction environment. *GCH*, 2014. 3

[17] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *ICCV*, 2015. 2, 5, 6

[18] 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 *CVPR*, 2020. 2, 3

[19] Ming Gui, Johannes S Fischer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*, 2024. 3

[20] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *CVPR*, 2022. 6

[21] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023. 6

[22] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. *CVPR*, 2024. 2, 4, 5, 6, 7

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 4

[24] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *ICCV*, 2023. 3

[25] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 3

[26] Michal Jancosek and Tomas Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR*, 2011. 3

[27] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014. 2, 6

[28] Yurim Jeon, Hwichang Kim, and Seung-Woo Seo. Abcd: Attentive bilateral convolutional network for robust depth completion. *RA-L*, 2021. 3

[29] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *CVPR*, 2024. 2, 3, 4, 6

[30] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *CVPR*, 2021. 3

[31] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *CVPR*, 2023. 3

[32] Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. 2024. 3

[33] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: the urbanscene3d dataset. In *ECCV*, 2022. 2, 6

[34] Yuankai Lin, Tao Cheng, Qi Zhong, Wending Zhou, and Hua Yang. Dynamic spatial propagation network for depth completion. In *AAAI*, 2022. 3

[35] Fei Liu, Zihao Lu, and Xianke Lin. Vision-based environmental perception for autonomous driving. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 2022. 3

[36] Xin Liu, Xiaofei Shao, Bo Wang, Yali Li, and Shengjin Wang. Graphcspn: Geometry-aware depth completion via dynamic gcns. In *ECCV*, 2022. 3

[37] Chen Long, Wenxiao Zhang, Zhe Chen, Haiping Wang, Yuan Liu, Peiling Tong, Zhen Cao, Zhen Dong, and Bisheng Yang. Sparsedc: Depth completion from sparse and non-uniform inputs. *Information Fusion*, 2024. 3

[38] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM TOG*, 2020. 3

[39] Fabian Märkert, Martin Sunkel, Anselm Haselhoff, and Stefan Rudolph. Segmentation-guided domain adaptation for efficient depth completion. *arXiv preprint arXiv:2210.09213*, 2022. 3

[40] Paul Merrell, Amir Akbarzadeh, Liang Wang, Philippos Mordohai, Jan-Michael Frahm, Ruigang Yang, David Nistér, and Marc Pollefeys. Real-time visibility-based fusion of depth maps. In *ICCV*, 2007. 2, 5

[41] Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Guided depth super-resolution by deep anisotropic diffusion. In *CVPR*, 2023. 3

[42] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011. 2, 4, 5

[43] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 2003. 5

[44] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *arXiv preprint arXiv:1909.05483*, 2019. 6

[45] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *CVPR*, 2019. 3

[46] 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 TPAMI*, 2020. 3

[47] 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 *ICCV*, 2021. 6

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3, 4, 5

[49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 4

[50] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *ECCV*, 2022. 2, 4

[51] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 3

[52] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 5

[53] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2, 3

[54] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2, 3

[55] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. *arXiv preprint arXiv:2406.01493*, 2024. 3

[56] Robust Multiview Stereopsis. Accurate, dense, and robust multiview stereopsis. *IEEE TPAMI*, 2010. 2, 3

[57] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 6

[58] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 5

[59] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *CVPR*, 2021. 2, 3

[60] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2, 6

[61] Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. Dn-splatter: Depth and normal priors for gaussian splatting and meshing. *arXiv preprint arXiv:2403.17822*, 2024. 3

[62] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 5

[63] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2, 3

[64] Xiaofeng Wang, Zheng Zhu, Guan Huang, Fangbo Qin, Yun Ye, Yijia He, Xu Chi, and Xingang Wang. Mvster: Epipolar transformer for efficient multi-view stereo. In *ECCV*, 2022. 2, 3

[65] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *ICCV*, 2021. 3

[66] Yuqun Wu, Jae Yong Lee, and Derek Hoiem. Sparse spn: Depth completion from sparse keypoints. *arXiv preprint arXiv:2212.00987*, 2022. 3

[67] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *CVPR*, 2023. 2, 3

[68] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. Depth completion from sparse lidar data with depth-normal constraints. In *ICCV*, 2019. 3

[69] Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024. 3

[70] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *CVPR*, 2024. 2, 3

[71] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 2024. 2, 3

[72] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *CVPR*, 2022. 2, 3

[73] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 2, 3, 4

[74] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *CVPR*, 2019. 2, 3

[75] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 2021. 2, 3

[76] 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 *ICCV*, 2023. 3

[77] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *NeurIPS*, 2022. 2, 3, 6

[78] W Yuan, X Gu, Z Dai, S Zhu, and P Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv preprint arXiv:2203.01502*, 2022. 3

[79] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *CVPR*, 2023. 3

[80] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. Geomvsnet: Learning multi-view stereo with geometry perception. In *CVPR*, 2023. 4

[81] Shanshan Zhao, Mingming Gong, Huan Fu, and Dacheng Tao. Adaptive context-aware multi-modal network for depth completion. *IEEE TIP*, 2021. 3