

Learning-based Monocular Depth Estimation for Photogrammetric 3D Reconstruction

Chunyu Dou^{1,2}, Yifei Yu^{1,*}, Xin Wang¹, Zongqian Zhan¹

¹ School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China

² School of Geography, Nanjing Normal University, Nanjing 210023, China
chunyu_dou@163.com, yfyu2020@whu.edu.cn, (xwang, zqzhan)@sgg.whu.edu.cn

Keywords: Monocular Depth Estimation, 3D reconstruction, Vision Transformer, Diffusion Models, Photogrammetric Images

Abstract

Monocular depth estimation (MDE) infers depth from a single image, offering significant advantages in computational efficiency and memory consumption compared to conventional Multi-View Stereo (MVS) methods. However, most MDE methods suffer from poor multi-view geometric consistency, which limits their application to photogrammetric 3D reconstruction. To address this issue, this paper employs sparse point clouds of Structure-from-Motion (SfM) as extra geometric constraints and proposes a framework that achieves photogrammetric 3D reconstruction using off-the-shelf learning-based MDE models without the need for additional fine-tuning. Specifically, when SfM priors are available during inference, globally geometrically consistent depth maps can be directly predicted. Otherwise, the estimated monocular depths are aligned to a consistent scale using SfM results via a post-correction step. The resulting depth maps are then fused using a truncated signed distance function (TSDF) to generate dense 3D reconstructions. Experiments on photogrammetric datasets demonstrate that the proposed framework effectively improves geometric consistency across depth maps and enables high-quality scene reconstruction. In addition, we systematically analyze the impact of key parameters in depth inference and fusion, including depth map resolution, voxel size, denoising steps, and ensemble size, on reconstruction performance, and further explore the potential of MDE for photogrammetric 3D reconstruction.

1. Introduction

Multi-view stereo (MVS), which aims to generate dense 3D points from images with known camera poses, plays a vital role in photogrammetric 3D reconstruction and has long been a research hotspot in the field of photogrammetry and computer vision. Conventional MVS methods (Furukawa and Ponce, 2009; Poggi and Mattoccia, 2016) perform accurate depth estimation by evaluating multiple depth hypotheses under multi-view geometric and photometric consistency constraints. However, the exhaustive hypothesis search and repeated multi-view consistency computation lead to substantial computational redundancy and memory overhead, limiting their efficiency when applied to large-scale photogrammetric datasets. To alleviate these limitations, learning-based MVS methods (Yao et al., 2018; Gu et al., 2020; Wang et al., 2021) have been investigated as a more efficient alternative, achieving strong robustness while reducing the reliance on explicit hypothesis enumeration. By implicitly learning disparity relationships and geometric consistency among multiple views via deep neural networks, these methods yield reliable depth maps and reconstructions even under challenging conditions such as various lighting or poor textures. Nevertheless, most of these approaches rely on 3D cost volume construction and 3D convolution, whose computational and memory complexity grows rapidly with increasing image resolution, thereby constraining both reconstruction resolution and overall efficiency.

In contrast, monocular depth estimation (MDE) predicts dense depth maps directly from a single image without relying on multi-view geometry, achieving lower computational cost and faster inference speed (Eigen et al., 2014; Ricci et al., 2018). Recently, both Vision Transformer (ViT)-based discriminative models (Yang et al., 2024; Xu et al., 2025) and diffusion-based

generative models (Duan et al., 2024; Ke et al., 2025) have demonstrated excellent performance in MDE tasks, particularly in terms of zero-shot generalization and high-quality depth prediction, offering a new perspective for depth estimation. However, these methods normally infer a well-looking and smooth depth solely for each image with internally inconsistent relative scales, and the corresponding multi-view geometric consistency is extremely poor, which results in scale ambiguity and geometric distortion when applied to multi-view 3D reconstruction.

To address these issues, recent studies (Guo et al., 2025; Viola et al., 2025) have attempted to incorporate sparse geometric priors obtained from Structure-from-Motion (SfM) (Schoenberger and Frahm, 2016) into MDE networks, aiming to improve multi-view geometric consistency while maintaining the computational efficiency of depth estimation. Inspired by these advancements, this paper proposes a framework for dense 3D reconstruction based on off-the-shelf learning-based MDE methods, which is applicable regardless of whether SfM geometric constraints are incorporated during inference. Specifically, SfM is employed to obtain sparse geometry before monocular depth inference. If SfM priors are available during inference, geometrically consistent depth maps are directly predicted and used for reconstruction. Otherwise, the predicted monocular depths are aligned to a consistent scale using SfM results via a post-correction step. Finally, all corrected depth maps are fused using TSDF for the dense 3D reconstruction. The main contributions of this work are threefold:

- A 3D reconstruction framework that integrates off-the-shelf MDE methods is proposed, which is applicable both with and without SfM geometric priors during depth inference;
- A comprehensive analysis of key parameters in depth inference and fusion is presented, including depth map resolution, voxel size, denoising steps, and ensemble inference;

* Corresponding author

- Extensive experiments are conducted on several photogrammetric datasets to validate the feasibility of the proposed framework and to explore the potential of learning-based monocular depth estimation methods for photogrammetric 3D dense reconstruction.

2. Related Work

In this section, three related topics are reviewed, including the conventional and learning-based Multi-view Stereo(MVS) methods, monocular depth estimation(MDE), as well as the feed-forward 3D reconstruction methods.

2.1 Multi-view Stereo

Multi-view stereo (MVS) is an important component of photogrammetric 3D reconstruction. Traditional MVS methods (Furukawa and Ponce, 2009; Poggi and Mattoccia, 2016; Schönberger et al., 2016) primarily rely on geometric and photometric consistency to perform dense matching across multiple views to estimate depth maps, which are subsequently fused for scene reconstruction. However, because of their limited feature representation capability, these methods often struggle to produce high-quality reconstructions in challenging scenes. Moreover, pixel-wise dense matching introduces substantial computational overhead, causing the cost of traditional MVS to increase rapidly with the number of input images and their resolution.

With the development of deep learning, learning-based MVS methods have attracted significant attention. MVNet (Yao et al., 2018) first introduced end-to-end learning into MVS by constructing a 3D cost volume and performing geometric regularization using 3D convolutional neural networks. Although it improves robustness in poor-texture regions, the heavy memory consumption of 3D convolution limits the reconstruction resolution. Subsequently, Gu et al. (2020) adopts a cascaded cost-volume strategy to incrementally refine depth estimation and improve efficiency. In addition, PatchMatchNet (Wang et al., 2021) integrates the classical PatchMatch sampling mechanism into neural networks, reducing memory consumption through iterative propagation and random search.

Despite these improvements, learning-based MVS methods still face several limitations. First, most methods rely on high-dimensional cost volumes, which leads to large memory consumption and limited scalability to large-scale scenes. Second, these approaches often depend on large amounts of ground-truth depth data for training, which restricts their generalization ability in unseen environments.

2.2 Monocular Depth Estimation

Monocular Depth Estimation (MDE) aims to infer scene depth from a single image. Unlike MVS, MDE only leverages semantic and geometric cues within a single image, significantly reducing computational complexity. Early MDE research (Eigen et al., 2014; Ricci et al., 2018) mainly employed convolutional neural networks (CNNs) with encoder-decoder architectures to perform end-to-end depth regression. However, due to their limited receptive field, CNN-based methods struggle to capture global scene structure in complex environments.

To address this issue, Vision Transformers (ViT) have been introduced into MDE tasks (Bhat et al., 2021; Ranftl et al., 2021;

Xu et al., 2025). Their self-attention mechanism enables global feature modeling, improving the estimation of overall scene layout and object relationships. Recently, DepthAnythingV2 (Yang et al., 2024) uses a large-scale visual pre-trained model, DINOv2 (Oquab et al., 2023) to achieve strong zero-shot generalization capability. In parallel, diffusion-based methods formulate depth estimation as a denoising process, progressively refining depth maps under image guidance (Duan et al., 2024; Ke et al., 2025). Although these diffusion models can generate high-resolution depth maps with rich details, their multi-step iterative inference typically incurs higher computational cost compared to ViT-based architectures.

Despite these advancements, MDE methods usually produce visually smooth depth predictions with inconsistent relative scales, resulting in poor geometric consistency across multiple views, which limits their direct application to dense multi-view reconstruction. To address this issue, recent studies such as Murre (Guo et al., 2025), Marigold-DC (Viola et al., 2025), and Prior-DepthAnything (Wang et al., 2025b) attempt to integrate sparse geometric priors from SfM into MDE networks. However, these approaches often require incorporating extra network modules and fine-tuning, which impose additional demands on computational resources and training data, and may even compromise the generalization ability of the models. In contrast, applying post-correction to align depth predictions from off-the-shelf MDE methods provides a practical and efficient alternative with strong applicability.

2.3 Feed-forward 3D Reconstruction Methods

More recently, feed-forward 3D reconstruction methods have attracted considerable attention due to their strong end-to-end reconstruction capability. These approaches can directly recover dense geometric structures from images in a single forward pass, without relying on traditional multi-stage processing pipelines.

DUST3R (Wang et al., 2024) is the first fully end-to-end reconstruction framework, which pioneeringly formulates pairwise reconstruction as a point map regression task. Subsequently, VGGT further improves the network architecture to enable the reconstruction of hundreds of images in a single forward pass, simultaneously predicting point maps, depth maps, camera parameters, and 3D point tracks. Building upon this work, subsequent methods have further improved robustness and flexibility. For instance, π^3 (Wang et al., 2025a) employs a fully permutation-equivariant architecture to enhance robustness, while MapAnything (Keetha et al., 2025) further scales to larger numbers of input images and supports the incorporation of additional geometric priors, such as camera poses and depth maps.

Although these feed-forward methods perform well in many challenging scenes and offer a fast inference speed, their high computational complexity and heavy GPU memory requirements limit their scalability to large-scale, high-resolution datasets. In contrast, the frame-wise inference paradigm of MDE methods makes them more suitable for processing large-scale image datasets.

3. Methodology

In this section, a brief overview of two representative learning-based MDE frameworks is first provided: the ViT-based discriminative MDE and the diffusion-based generative MDE.

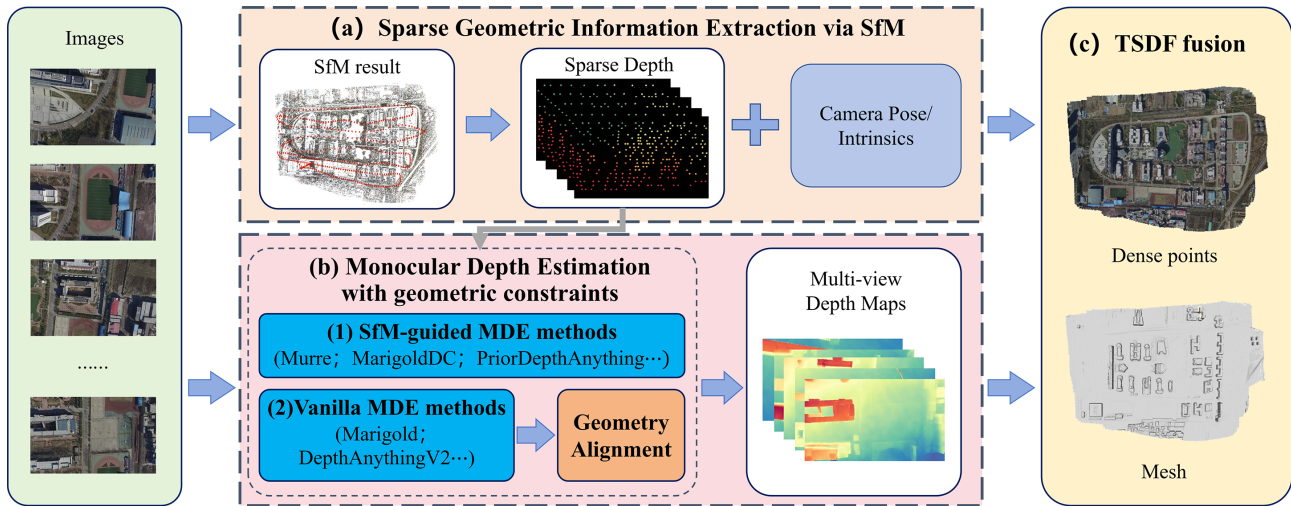


Figure 1. Workflow of 3D reconstruction framework based on monocular depth estimation.

This is followed by a detailed description of the proposed MDE-based reconstruction framework, which consists of three main processes: 1) sparse geometric information extraction via SfM; 2) monocular depth estimation with geometric constraints; 3) 3D reconstruction via TSDF fusion.

3.1 Learning-Based Monocular Depth Estimation

The goal of monocular depth estimation is to predict a pixel-wise dense depth map D from a single image I . Recent learning-based MDE methods can be roughly divided into two categories: ViT-based discriminative MDE and Diffusion-based generative MDE.

3.1.1 ViT-based discriminative MDE: ViT-based discriminative MDE (Yang et al., 2024), typically employs Vision Transformers as the backbone network, treating depth inference as a feature regression task. At the input stage, the ViT divides the image into fixed-size patches and encodes them into high-dimensional visual tokens. Through the self-attention mechanism, each token interacts with other tokens across the entire feature space, enabling effective capture of long-range dependencies and global contextual information within the image.

In recent years, many approaches have employed large-scale visual pre-trained models (e.g., DINOv2) as the feature encoder to obtain rich semantic and geometric representations. These multi-scale intermediate features from different encoder stages are incorporated into a decoder, enabling the aggregation of contextual information at multiple resolutions. The decoder progressively recovers spatial details via upsampling and skip connections, yielding refined features that are finally mapped by a lightweight regression head into a dense depth map.

3.1.2 Diffusion-based generative MDE: Diffusion-based generative MDE (Ke et al., 2025) formulates depth estimation as a latent conditional diffusion process (Rombach et al., 2022), in which depth is gradually recovered from random noise through iterative denoising. During the forward diffusion process, starting from a clean depth map d_0 , Gaussian noise is progressively injected over T steps to generate a sequence of noisy samples $d_1, d_2, \dots, d_T$. According to the Markov chain property, the noise depth at step t can be expressed as Eq. (1):

$$d_t = \sqrt{\bar{\alpha}_t} d_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (1)$$

where $\bar{\alpha}_t$ denotes the predefined variance schedule, defined as $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. When $t \rightarrow T$, d_T approaches isotropic Gaussian noise. During the reverse denoising process, a pre-trained denoising network ϵ_θ (typically based on a U-Net) is used to gradually recover the clean depth map d_0 from the noisy sample d_T under the guidance of the input image I . At each iteration step t , the network predicts the residual noise according to Eq. (2):

$$\epsilon_t = \epsilon_\theta(d_t, E(I), t) \quad (2)$$

where $E(I)$ represents the feature encoding of the image condition. By iteratively applying the mean prediction formulation (e.g., the DDIM sampler), the model finally generates depth maps with rich structural details and consistent geometry. Diffusion-based methods benefit from their ability to leverage learned visual priors, enabling robust depth estimation in occluded or poor-texture regions.

3.2 3D Reconstruction Framework Based on Monocular Depth Estimation

Fig. 1 illustrates the workflow of the proposed 3D reconstruction framework based on monocular depth estimation. The pipeline integrates sparse geometric information obtained from Structure-from-Motion (SfM) with monocular depth estimation to generate scale-consistent depth maps, which are subsequently fused using the TSDF to produce dense point clouds and surface meshes.

3.2.1 Sparse Geometric Information Extraction via SfM: First, a widely recognized SfM method, COLMAP (Schonberger and Frahm, 2016) is applied to obtain sparse geometric information from the input images. This step estimates the camera intrinsic matrix K , the camera poses $[R_i \mid t_i]$ for each image, and the sparse 3D point cloud P of the scene. These sparse 3D points are then reprojected onto the image plane to generate sparse depth maps $D_{\text{sparse}} \in \mathbb{R}^{H \times W}$, where the metric depth values are filled only at pixels corresponding to the SfM observations (Hartley and Zisserman, 2003). The resulting sparse depth maps provide reliable geometric references for subsequent scale recovery in monocular depth estimation and for TSDF fusion.

3.2.2 Monocular Depth Estimation with Geometric Constraints: Monocular depth estimation methods usually rely

only on semantic information from a single image, lacking explicit geometric constraints between multiple views. As a result, when applied to large-scale or complex scenes, the predicted depths from different views often suffer from severe scale drift and multi-view geometric inconsistency. To address this issue, this work investigates two technical paths for obtaining geometric consistent depth maps.

(1) In-network Guidance with SfM Priors: The core idea of this approach is to directly integrate the sparse depth map D_{sparse} obtained in Section 3.2.1 as a reliable geometric constraint for the depth inference process. Two guidance strategies are employed, depending on the model architecture.

Guidance in diffusion-based models—Murre (Guo et al., 2025), Marigold-DC (Viola et al., 2025): In diffusion-based methods, the explicit geometric constraints provided by D_{sparse} are used to guide the denoising process, ensuring the depths are generated in a consistent scale. Specifically, take Murre and Marigold-DC as examples.

In Murre, the sparse depth map is first densified using k-nearest neighbors (KNN) interpolation. A distance-weight map is also computed to represent the Euclidean distance of each pixel to its nearest SfM point, allowing the model to distinguish between real observations and interpolated depths, which suppresses interpolation noise. The densified depth map and the distance map are then provided as additional conditions to guide the iterative denoising process, introducing geometric constraints during depth inference.

In contrast, Marigold-DC adopts a test-time optimization strategy. During each denoising iteration, the Tweedie formula is used to decode the current latent variable into an intermediate non-metric depth map d_t . Since the model predicts an affine-invariant relative depth, sparse SfM points are used for linear initialization to estimate a scale factor a and offset b , yielding a metric depth estimate $d_t^{(m)} = ad_t + b$. A loss function L is then computed between the metric depth $d_t^{(m)}$ and the sparse SfM depths. The resulting error is backpropagated to the latent variables, progressively improving geometric consistency during denoising.

Guidance in ViT-based models—Prior-DepthAnything (Wang et al., 2025b): In this strategy, the sparse depth map D_{sparse} is first used to linearly transform the relative depth predicted by an initial MDE model into a coarse metric space. The original RGB image, the pre-aligned depth map, and a mask indicating SfM observation locations are then jointly fed into a conditional MDE model for refinement. Benefiting from the global attention mechanism, the model effectively propagates sparse metric constraints across the entire image, significantly enhancing geometric consistency.

(2) Explicit Geometric Post-Correction Based on Robust Regression: A fundamental limitation of most monocular depth estimation models lies in their intrinsic scale ambiguity, as they primarily predict relative depth only to an unknown scale, e.g., DepthAnythingV2 (Yang et al., 2024), Marigold (Ke et al., 2025), and PixelPerfectDepth (Xu et al., 2025). Consequently, inconsistent scaling across multiple views leads to significant scale drift, which may result in reconstruction failure in multi-view reconstruction tasks, as shown in Fig. 2.

To address this problem, an explicit geometric post-correction strategy based on sparse SfM constraints is proposed. The goal

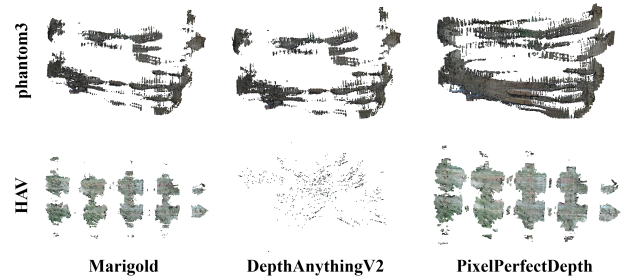


Figure 2. Reconstruction results of vanilla MDE methods.

is to align monocular depth predictions into a unified scale through a lightweight transformation, without modifying the pre-trained model.

First, during the reprojection process described in Section 3.2.1, only stable 3D points observed in a sufficient number of views are retained to ensure reliable geometric constraints. Meanwhile, to handle potential occlusion during reprojection, a Z-buffer strategy is implemented to resolve depth conflicts at the same pixel location.

To avoid the influence of depth discontinuities, local depth variation is estimated using morphological dilation and erosion $\Delta d = d_{\text{dilate}} - d_{\text{erode}}$. Pixels satisfying $\Delta d/d > \tau$ are identified as depth discontinuities and excluded from regression, preventing boundary artifacts from biasing scale estimation.

After obtaining reliable pixel correspondences, a regression relationship is established between the predicted depth d_{pred} and the sparse depth from SfM d_{sfm} . For depth predictions in linear space, an affine transformation is assumed:

$$d_{\text{sfm}} = a \cdot d_{\text{pred}} + b \quad (3)$$

where a and b represent the scale factor and offset. For predictions better modeled in disparity space, regression is instead performed in the inverse depth domain:

$$\frac{1}{d_{\text{sfm}}} = a \cdot \frac{1}{d_{\text{pred}}} + b \quad (4)$$

The parameters are estimated using RANSAC (Random Sample Consensus), which can effectively handle outliers and suppress the influence of incorrect correspondences. After obtaining the optimal regression model, the resulting transformation is applied to the full depth map to obtain scale-consistent predictions.

Through this process, sparse geometric information from SfM is used to align monocular depth predictions to unified metric scale without modifying the original network architecture, thereby significantly improving multi-view depth consistency while preserving the efficiency and generalization capability of the original MDE model.

3.2.3 3D Reconstruction via TSDF Fusion: After obtaining metric depth maps for all images, TSDF (truncated signed distance function) fusion is employed for volumetric integration (Curless and Levoy, 1996; Newcombe et al., 2011). Using the camera parameters estimated by SfM, depth observations from multiple views are projected into a unified 3D voxel grid.

Res.	Vox.	Murre					Marigold-DC				
		C2C↓	Rough↓	T_{infer} ↓	T_{fuse} ↓	Mem.↓	C2C↓	Rough↓	T_{infer} ↓	T_{fuse} ↓	Mem.↓
512	5	0.256	0.021	1.43	3.07	24.57	Out of Memory				
512	10	0.262	0.045	1.43	1.17	6.62	0.217	0.045	44.90	2.08	26.45
512	20	0.252	0.095	1.43	0.91	4.93	0.193	0.096	44.90	0.68	7.60
512	30	0.237	0.142	1.43	0.37	4.53	0.199	0.146	44.90	0.92	6.54
1024	5	0.227	0.021	5.81	1.69	26.02	Out of Memory				
1024	10	0.227	0.047	5.81	0.87	6.85	0.240	0.044	51.52	2.40	26.32
1024	20	0.214	0.099	5.81	0.59	5.10	0.211	0.094	51.52	0.62	7.55
1024	30	0.217	0.150	5.81	0.40	4.75	0.212	0.143	51.52	0.78	6.51
1536	5	0.236	0.021	16.74	1.79	27.25	Out of Memory				
1536	10	0.238	0.046	16.74	1.09	7.34	0.214	0.044	106.26	2.06	26.12
1536	20	0.213	0.097	16.74	0.78	5.54	0.210	0.093	106.26	0.62	7.51
1536	30	0.216	0.148	16.74	0.40	5.20	0.211	0.142	106.26	0.75	6.49

Table 1. Performance comparison of Murre and Marigold-DC under different depth resolutions and voxel parameters.

During fusion, the signed distance value and corresponding weight of each voxel are incrementally updated. When the TSDF fusion is complete, the marching cubes algorithm is applied to extract the zero-crossing isosurface from the TSDF volume, which represents the scene surface. The extracted isosurface is then used to generate dense point clouds and mesh models.

4. Experiments

To systematically evaluate the performance of the proposed MDE-based 3D reconstruction framework on photogrammetry datasets, experiments are designed from two key perspectives. First, the influence of different parameters on reconstruction is investigated to determine suitable configurations for the framework. Based on these settings, reconstruction is conducted with multiple MDE methods to verify the effectiveness of the proposed framework.

Specifically, Section 4.2 examines the impact of depth map resolution and voxel size on the quality of TSDF-based reconstruction. Section 4.3 investigates how the number of denoising steps and ensemble size in diffusion-based methods affect the reconstruction results. Finally, in Section 4.4, guided by the conclusion of Sections 4.2 and 4.3, six state-of-the-art MDE methods (Murre, Marigold, Marigold-DC, Prior-DepthAnything, PixelPerfectDepth, DepthAnythingV2) are employed for reconstruction, and their performance is analyzed within the proposed framework.

4.1 Datasets

To comprehensively evaluate the 3D reconstruction performance of the proposed framework under different photogrammetric conditions, several typical datasets are used in the experiments. (1) Low-altitude UAV photogrammetry datasets: CAS (Du et al., 2025), SHHY, and GauU (Xiong et al., 2024). For the CAS dataset, two scenes are selected, including phantom3 and ground. The GauU dataset includes two large-scale scenes, HAV and SMBU, which are used to assess the reconstruction capability of the proposed framework in large-scale UAV scenarios. The SHHY dataset contains complex urban structures with high image resolution, and is used to evaluate reconstruction performance on high-resolution UAV images. (2) Close-range photogrammetry dataset: the Caffe in FBK dataset

(Poiesi et al., 2017) is mainly used to evaluate the reconstruction performance in small-scale close-range environments. Examples and detailed information of these datasets are presented in Fig. 3 and Table 2.

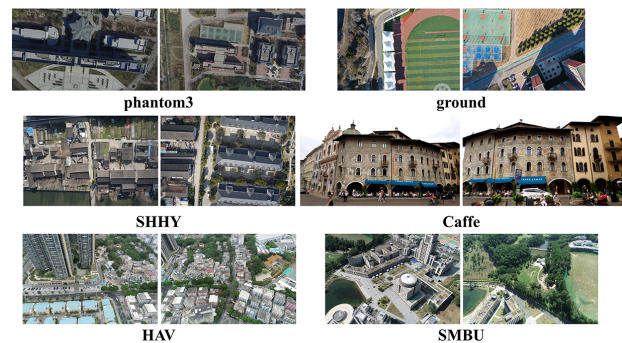


Figure 3. Sample images of experimental datasets.

Dataset	Resolution (px)	Num. Images
phantom3	1870×1049	467
ground	5472×3648	295
SHHY	6000×4000	770
Caffe	1920×1080	87
HAV	5467×3635	424
SMBU	5468×3636	563

Table 2. The resolution and number of images of datasets used in experiments

Regarding the ground-truth reference, CAS and SHHY datasets use high-density point clouds reconstructed by Agisoft Metashape as reference model. The FBK dataset directly adopts the provided LiDAR point clouds. For the GauU dataset, official LiDAR point clouds are used as the reference, while the SfM camera poses are reconstructed using COLMAP and further aligned to LiDAR reference to ensure scale consistency during evaluation.

All experiments and evaluations in this work are conducted on a workstation equipped with an Intel(R) Xeon(R) Gold 6133 CPU and a single NVIDIA GeForce RTX 4090 GPU (24 GB VRAM). The system operates on Ubuntu 22.04.3 LTS with 512 GB of physical RAM, ensuring sufficient computational resources for high-resolution 3D reconstruction and large-scale

TSDF fusion.

4.2 Parameter Analysis of Voxel Size and Depth Map Resolution

This section evaluates the influence of depth map resolution (Res., in pixels) and voxel size (Vox., calculated as Voxel Size = Voxel / 512 m) on the reconstruction accuracy during TSDF fusion. The experiments are conducted on the FBK-Caffe scene using Murre and Marigold-DC as examples, as shown in Tab. 1. Here, C2C (in meters) denotes the mean Cloud-to-Cloud distance between the reconstructed and reference point clouds. Rough. (in meters) measures the mean surface roughness of the reconstructed model. T_{infer} and T_{fusion} (in seconds) denote the average inference time and fusion time per image, respectively, while Mem. (in GB) represents the peak memory consumption during the fusion process.

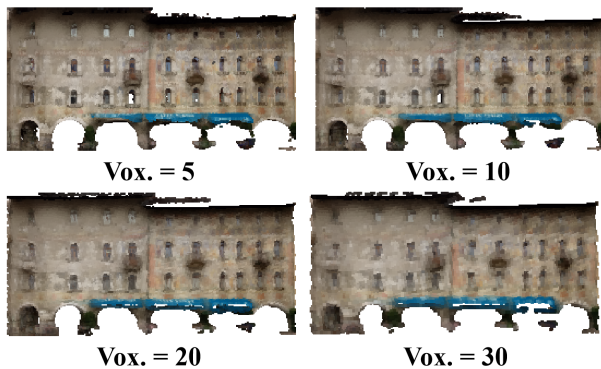


Figure 4. The reconstruction results of different voxel size on Caffe dataset

Under a fixed depth map resolution, increasing the voxel size leads to a slight decrease in C2C error. This is likely due to the stronger smoothing effect introduced by larger voxels, which suppresses noise in the reconstructed geometry, resulting in marginal improvements in numerical accuracy. However, this improvement in C2C error does not necessarily reflect an actual increase in geometric fidelity. In contrast, surface roughness increases significantly with larger voxel sizes, as shown in Fig. 4, indicating that the ability to preserve geometric details is substantially degraded. Moreover, reducing the voxel size results in a significant increase in both fusion time and memory consumption.

When the voxel size is fixed, varying the depth map resolution has a relatively limited impact on both C2C error and surface roughness, with variations generally within 2 cm. Although higher-resolution depth maps provide denser spatial observations, the TSDF fusion mechanism inherently aggregates all depth observations falling within the same voxel through weighted averaging. Consequently, the potential geometric gains from increased resolution are mitigated by this localized fusion. In terms of computational cost, increasing the depth map resolution leads to a significant increase in inference time, while having minimal effect on fusion time.

Overall, both voxel size and depth map resolution have limited influence on C2C accuracy, whereas voxel size plays a dominant role in determining surface quality. In particular, larger voxels lead to smoother but less detailed surfaces, while smaller voxels preserve finer geometry at the cost of substantially higher computational and memory requirements. Similarly, excessively high depth map resolution increases inference cost

without proportional gains in reconstruction quality. Therefore, considering the trade-off between reconstruction quality and computational efficiency, Res. = 1024 and Vox. = 10 are adopted as the default parameter settings in subsequent experiments.

4.3 Parameter Analysis in Diffusion-based Depth Estimation

This section analyzes the influence of denoising steps (N_{step}) and ensemble size (N_{ensemble}) in diffusion models on the final 3D reconstruction results. Experiments are conducted using three diffusion-based models: Murre, Marigold-DC, and Marigold. Results are presented for the ground scene of the CAS dataset. Tab. 3, 4, and 5 report inference time, fusion time, C2C error and mean surface roughness. It is worth noting that both C2C and Rough. are computed in a normalized coordinate system for this dataset, and thus do not reflect absolute real-world scale.

N_{step}	N_{ens}	$T_{\text{infer}} \downarrow$	$T_{\text{fusion}} \downarrow$	C2C $\downarrow$	Rough. $\downarrow$
1	1	1.24	1.05	14.570	0.333
2	1	1.29	0.59	1.440	0.557
4	1	1.44	0.53	1.574	0.566
4	5	3.24	0.53	1.594	0.566
4	10	4.87	0.54	1.560	0.568
10	1	1.64	0.52	1.606	0.569
10	5	6.94	0.48	1.694	0.570
20	5	6.63	0.52	1.636	0.573

Table 3. Sensitivity analysis for the Murre model

N_{step}	N_{ens}	$T_{\text{infer}} \downarrow$	$T_{\text{fusion}} \downarrow$	C2C $\downarrow$	Rough. $\downarrow$
10	1	2.18	1.73	35.575	0.384
10	5	5.53	1.81	34.757	0.387
10	10	14.54	1.72	33.243	0.387
20	5	14.42	1.64	12.749	0.399
50	5	33.67	1.40	3.792	0.465

Table 4. Sensitivity analysis for the Marigold-DC model.

N_{step}	N_{ens}	$T_{\text{infer}} \downarrow$	$T_{\text{align}} \downarrow$	$T_{\text{fusion}} \downarrow$	C2C $\downarrow$	Rough. $\downarrow$
1	1	5.02	0.50	1.35	1.949	0.621
2	1	5.81	0.49	0.88	1.937	0.627
4	1	6.48	0.40	1.09	1.870	0.638
4	5	7.85	0.40	3.32	1.956	0.643
4	10	17.49	0.75	2.80	2.014	0.636
10	1	6.15	0.36	0.99	2.001	0.629

Table 5. Sensitivity analysis for the Marigold model.

Murre: The results show that the number of denoising steps N_{step} significantly affects reconstruction accuracy, while the influence of ensemble size N_{ensemble} is relatively limited. When N_{step} increases from 1 to 2, the C2C error decreases sharply, whereas surface roughness increases, indicating over-smoothing at extremely low denoising steps. In this regime, the model fails to recover both metric scale and essential geometric details, which results in severe depth prediction errors. As N_{step} further increases to a range of 4–20, both mean C2C and roughness stabilize at approximately 1.6, indicating that the model reaches a stable prediction state with relatively few sampling steps. This is mainly attributed to the sparse SfM geometric priors integrated in Murre, which facilitate rapid convergence

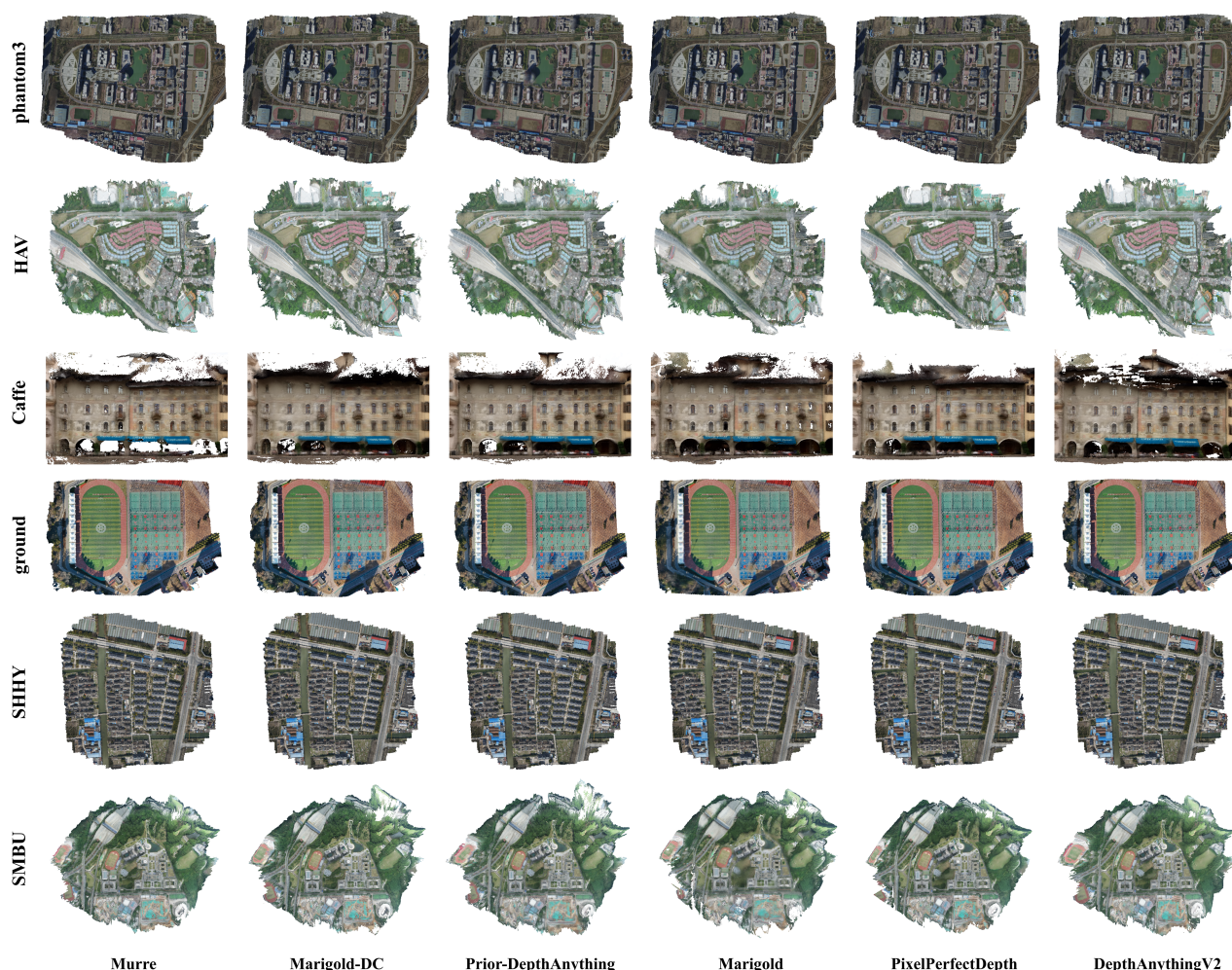


Figure 5. Dense reconstruction results of several monocular depth estimation methods

during the diffusion process. Increasing N_{ensemble} yields only marginal improvements in C2C but significantly increases inference time. Therefore, considering both efficiency and accuracy, $N_{\text{step}} = 2$ and $N_{\text{ensemble}} = 1$ achieve a good balance.

Marigold-DC: The experimental results show that the diffusion steps have a stronger influence on depth prediction accuracy. This method relies on test-time optimization in the latent space to recover depth scale, making it highly sensitive to the number of diffusion iterations. When N_{step} is small, the latent optimization process fails to sufficiently converge, leading to large reconstruction errors. As N_{step} increases, the error decreases significantly, accompanied by an increase in surface roughness, indicating that sufficient diffusion iterations are required for reliable scale recovery and alignment with SfM constraints. However, increasing N_{step} leads to substantially higher inference time, making it significantly slower than other compared methods. In the meanwhile, increasing N_{ensemble} provides limited improvement in accuracy. Therefore, considering computational efficiency, $N_{\text{step}} = 50$ and $N_{\text{ensemble}} = 5$ are adopted as the default configuration for reconstruction.

Marigold: The results show that stable depth prediction can be obtained with relatively few diffusion steps. When N_{step} increases from 1 to 4, the C2C error decreases from 1.949 to 1.870, while roughness increases from 0.621 to 0.638, indicating that a moderate increases in diffusion steps improve both reconstruction accuracy and geometric detail. However, when

N_{step} further increases to 10, the reconstruction error slightly increases, which suggests that excessive diffusion steps may introduce noise accumulation and degrade performance. Regarding ensemble inference, when $N_{\text{step}} = 4$, increasing N_{ensemble} from 1 to 5 or 10 does not significantly improve reconstruction accuracy but leads to additional computational consumption in alignment and fusion stages. Therefore, $N_{\text{step}} = 4$ and $N_{\text{ensemble}} = 1$ provide a reasonable trade-off between reconstruction accuracy and computational efficiency.

4.4 Reconstruction Results of the Proposed Framework

This section evaluates the performance of six state-of-the-art monocular depth estimation methods within the proposed 3D reconstruction framework on six scenes: phantom3, HAV, Caffe, ground, SHHY and SMBU. All these methods use the same TSDF fusion parameters (Res. = 1024, Vox. = 10). Reconstruction results of these MDE methods are shown in Fig. 5, quantitative evaluation results are presented in Tab. 6

To assess reconstruction quality, four metrics are adopted: the mean Cloud-to-Cloud error (C2C mean) between the reconstructed and reference point clouds, the 95th percentile of the C2C error distribution (C2C P95), the F-score computed using C2C mean as the distance threshold, and the mean surface roughness (Rough.). The visualization results of the mean C2C errors for phantom3, HAV and Caffe are shown in Fig. 6, 7, and 8. It should be noted all reconstructed point clouds are aligned

Dataset	MDE Methods	C2C Mean↓	C2C P95↓	F-Score↑	Rough.↓	T_{total} ↓	VRAM↓
phantom3 (1870×1049)	Murre	0.230	0.550	<u>0.895</u>	<u>0.078</u>	1556	9.56
	Marigold-DC	0.434	1.619	0.788	0.064	102284	7.51
	Prior-DepthAnything*	0.231	0.570	0.880	0.080	270	2.22
	Marigold	0.229	0.555	0.881	0.082	2179	<u>4.90</u>
	PixelPerfectDepth	0.224	0.538	0.900	0.084	<u>646</u>	5.61
	DepthAnythingV2*	<u>0.225</u>	<u>0.549</u>	<u>0.895</u>	<u>0.078</u>	738	6.17
HAV (5467×3635)	Murre	0.092	0.457	<u>0.979</u>	0.008	2307	11.19
	Marigold-DC	0.081	0.506	0.974	0.008	30521	8.54
	Prior-DepthAnything*	0.074	0.435	0.980	0.008	2017	11.86
	Marigold	0.083	0.508	0.882	0.008	3625	5.45
	PixelPerfectDepth	<u>0.069</u>	<u>0.408</u>	0.900	0.008	<u>2171</u>	5.68
	DepthAnythingV2*	0.064	0.391	0.913	0.008	2686	<u>5.48</u>
Caffe (1920×1080)	Murre	0.274	0.729	0.873	0.043	449	9.70
	Marigold-DC	<u>0.240</u>	<u>0.714</u>	<u>0.892</u>	<u>0.044</u>	4691	7.52
	Prior-DepthAnything*	0.193	0.681	0.924	0.049	240	2.22
	Marigold	0.279	0.720	0.875	<u>0.044</u>	201	<u>5.09</u>
	PixelPerfectDepth	0.272	<u>0.714</u>	0.879	0.045	<u>227</u>	5.61
	DepthAnythingV2*	0.257	0.727	0.879	0.045	287	6.17
ground (5472×3648)	Murre	1.438	4.042	0.752	0.268	895	10.88
	Marigold-DC	2.752	6.768	0.583	0.215	61538	8.63
	Prior-DepthAnything*	<u>1.354</u>	3.637	0.773	<u>0.268</u>	2616	11.90
	Marigold	2.088	7.924	0.684	0.314	2429	5.25
	PixelPerfectDepth	1.339	<u>3.974</u>	<u>0.759</u>	0.291	2795	5.68
	DepthAnythingV2*	1.483	4.386	0.723	0.285	<u>1987</u>	<u>5.48</u>
SHHY (6000×4000)	Murre	0.063	0.173	0.984	<u>0.050</u>	3993	10.87
	Marigold-DC	0.180	0.499	0.845	0.075	136658	8.80
	Prior-DepthAnything*	<u>0.067</u>	<u>0.184</u>	<u>0.978</u>	0.051	3022	14.00
	Marigold	0.113	0.306	0.904	0.055	8731	<u>5.27</u>
	PixelPerfectDepth	0.074	0.213	0.973	0.052	<u>6121</u>	5.68
	DepthAnythingV2*	0.073	0.211	0.972	0.048	<u>7097</u>	5.48
SMBU (5468×3636)	Murre	0.279	1.484	0.871	0.009	3263	8.56
	Marigold-DC	<u>0.297</u>	1.619	0.862	0.011	57607	8.55
	Prior-DepthAnything*	0.331	1.788	0.855	0.010	2984	11.86
	Marigold	0.279	1.387	0.859	0.009	6064	<u>5.45</u>
	PixelPerfectDepth	0.299	<u>1.575</u>	0.858	0.011	3192	5.68
	DepthAnythingV2*	0.298	1.622	<u>0.868</u>	0.011	<u>3083</u>	5.48

* ViT-based discriminative monocular depth estimation methods; others are diffusion-based generative methods.

Table 6. **Quantitative evaluation of various MDE methods across six datasets.** Mean C2C, C2C P95, F-score, surface roughness, total processing time T_{total} , and VRAM usage are reported. Best and second-best results are highlighted in **bold** and underlined, respectively. For each dataset, the first three methods incorporate SfM priors during inference, while the last three apply post-correction to predicted relative depths.

to the coordinate system of the reference data. For datasets with LiDAR ground truth (e.g., HAV, Caffe and SMBU), the C2C error and Roughness are expressed in meters, reflecting absolute physical scales. Conversely, for datasets without LiDAR measurements (e.g., phantom3, ground and SHHY), these metrics serve primarily as relative indicators for inter-method comparison. In addition, the total reconstruction time T_{total} and peak memory consumption (Mem.) of each MDE method are compared.

In terms of reconstruction accuracy, Murre, Prior-DepthAnything, PixelPerfectDepth, and DepthAnythingV2 achieve the best performance on different datasets, respectively, with relatively small differences observed among them on the same dataset. This indicates that both strategies—incorporating

SfM geometric priors during network inference (e.g., Murre and Prior-DepthAnything) and applying SfM-based post-correction to align predicted depths (e.g., PixelPerfectDepth and DepthAnythingV2) are effective in improving multi-view geometric consistency and enabling reliable 3D reconstruction, with comparable levels of geometric accuracy. In addition, the surface roughness of all six MDE methods is similar, suggesting that roughness is less sensitive to the choice of MDE method and is primarily governed by the voxel size in the TSDF fusion process.

For computational efficiency, ViT-based methods generally perform higher computational efficiency than diffusion-based approaches, especially for Prior-DepthAnything, which achieves the shortest total processing time across four datasets. This

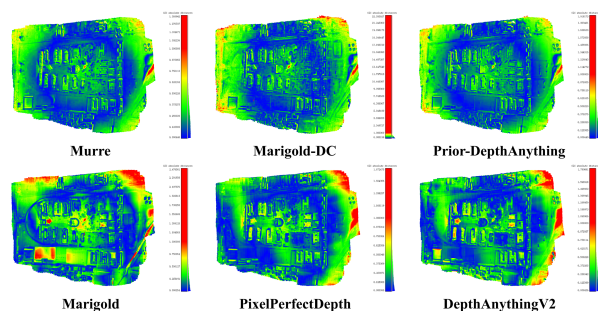


Figure 6. Cloud-to-cloud error comparison of several depth estimation models on phantom3 dataset

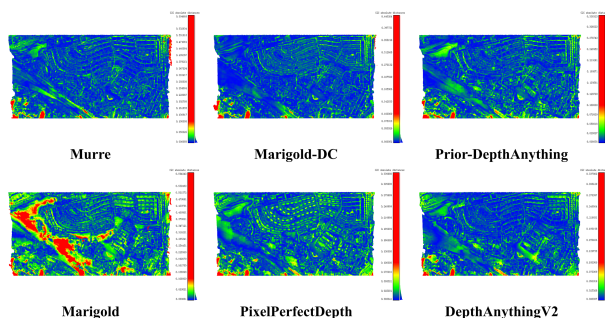


Figure 7. Cloud-to-cloud error comparison of several depth estimation models on HAV dataset

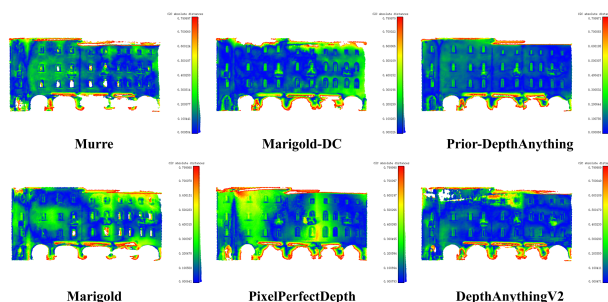


Figure 8. Cloud-to-cloud error comparison of several depth estimation models on Caffe dataset

is mainly because diffusion-based methods typically require multiple denoising steps and ensemble inference size to ensure prediction quality, resulting in longer inference time and higher computational cost. Notably, Murre demonstrates the highest efficiency among diffusion-based methods, as it effectively leverages SfM priors to achieve high-quality results with relatively few denoising steps and a small ensemble size.

For GPU memory consumption during inference, Prior-DepthAnything exhibits a clear advantage on low-resolution images but becomes less efficient than other methods when applied to high-resolution inputs. In general, methods that do not incorporate SfM priors during inference (e.g., Marigold, PixelPerfectDepth, and DepthAnythingV2) require less GPU memory than those that explicitly utilize SfM priors (e.g., Murre, Marigold-DC, and Prior-DepthAnything). Among them, Marigold consistently achieves the best or second-best memory efficiency across all evaluated datasets. This is because they do not process additional geometric information during inference, resulting in simpler network architectures and lower memory overhead.

Overall, the proposed MDE-based 3D reconstruction framework achieves reliable reconstruction performance across all six evaluated MDE methods. In particular, the proposed explicit post-correction strategy effectively improves multi-view geometric consistency for methods that lack geometric constraints during inference, while introducing only minimal computational overhead. Furthermore, under the proposed framework, ViT-based MDE methods achieve comparable reconstruction accuracy to diffusion-based approaches while offering significantly higher efficiency. Therefore, combining ViT-based MDE methods with the proposed post-correction strategy provides a favorable balance between reconstruction accuracy and computational efficiency.

5. Conclusion

This paper presents a novel 3D reconstruction framework that synergizes Monocular Depth Estimation (MDE) with Structure-from-Motion (SfM) geometric constraints for photogrammetric dense reconstruction. Central to our approach is an explicit geometric post-correction strategy that aligns the predicted monocular depth with the metric scale without requiring architectural modifications to pre-trained networks. This plug-and-play design effectively mitigates the multi-view geometric inconsistencies inherent in standard MDE pipelines.

Our extensive experiments reveal critical insights into the factors governing reconstruction quality. We observe that while both depth map resolution and voxel size influence the geometric fidelity of TSDF fusion, voxel size exerts a more dominant impact. Furthermore, for diffusion-based models, the number of denoising steps plays a pivotal role in ensuring scale convergence, particularly within optimization-based pipelines characterized by strong step dependency. While ensemble inference size effectively suppresses depth noise, we note diminishing returns in accuracy as ensemble size increases, accompanied by a linear rise in computational cost. Crucially, the proposed post-correction strategy demonstrates robust compatibility across both ViT-based and diffusion-based backbones, substantially enhancing multi-view geometric consistency. Notably, ViT-based models, when augmented with our strategy, offer a superior trade-off: they achieve high reconstruction accuracy with markedly higher computational efficiency compared to diffusion-based counterparts. This provides a more optimal balance between reconstruction fidelity and resource consumption.

Future work will explore more refined depth alignment strategies, such as patch-wise scale correction or multi-model regression, to further enhance geometric fidelity and robustness in complex, large-scale scenarios.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (42301507).

References

Bhat, S. F., Alhashim, I., Wonka, P., 2021. Adabins: Depth estimation using adaptive bins. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4009–4018.

- Curless, B., Levoy, M., 1996. A volumetric method for building complex models from range images. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 303–312.
- Du, B., Ye, H., Zhang, Y., Liao, X., 2025. A real-time orthophoto generation approach for uav based on deep learning visual feature. *IEEE Transactions on Geoscience and Remote Sensing*.
- Duan, Y., Guo, X., Zhu, Z., 2024. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. *European Conference on Computer Vision*, Springer, 432–449.
- Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27.
- Furukawa, Y., Ponce, J., 2009. Accurate, dense, and robust multi-view stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8), 1362–1376.
- Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P., 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2495–2504.
- Guo, H., Zhu, H., Peng, S., Lin, H., Yan, Y., Xie, T., Wang, W., Zhou, X., Bao, H., 2025. Multi-view reconstruction via sfm-guided monocular depth estimation. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5272–5282.
- Hartley, R., Zisserman, A., 2003. *Multiple view geometry in computer vision*. Cambridge university press.
- Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A., Schindler, K., 2025. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N. et al., 2025. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*.
- Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A., 2011. Kinectfusion: Real-time dense surface mapping and tracking. *2011 10th IEEE international symposium on mixed and augmented reality*, Ieee, 127–136.
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. et al., 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
- Poggi, M., Mattoccia, S., 2016. Learning a general-purpose confidence measure based on o (1) features and a smarter aggregation strategy for semi global matching. *2016 Fourth international conference on 3D vision (3DV)*, IEEE, 509–518.
- Poiesi, F., Locher, A., Chippendale, P., Nocerino, E., Remondino, F., Van Gool, L., 2017. Cloud-based collaborative 3d reconstruction using smartphones. *Proceedings of the 14th European Conference on Visual Media Production (CVMP 2017)*, 1–9.
- Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.
- Ricci, E., Ouyang, W., Wang, X., Sebe, N. et al., 2018. Monocular depth estimation using multi-scale continuous crfs as sequential deep networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(6), 1426–1440.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.
- Schönberger, J. L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise view selection for unstructured multi-view stereo. *European conference on computer vision*, Springer, 501–518.
- Viola, M., Qu, K., Metzger, N., Ke, B., Becker, A., Schindler, K., Obukhov, A., 2025. Marigold-dc: Zero-shot monocular depth completion with guided diffusion. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5359–5370.
- Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M., 2021. Patchmatchnet: Learned multi-view patchmatch stereo. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14194–14203.
- Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J., 2024. Dust3r: Geometric 3d vision made easy. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20697–20709.
- Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T., 2025a. π^3 : Permutation-Equivariant Visual Geometry Learning. *arXiv preprint arXiv:2507.13347*.
- Wang, Z., Chen, S., Yang, L., Wang, J., Zhang, Z., Zhao, H., Zhao, Z., 2025b. Depth anything with any prior. *arXiv preprint arXiv:2505.10565*.
- Xiong, B., Zheng, N., Liu, J., Li, Z., 2024. Gauu-scene v2: Assessing the reliability of image-based metrics with expansive lidar image dataset using 3dgs and nerf. *arXiv preprint arXiv:2404.04880*.
- Xu, G., Lin, H., Luo, H., Wang, X., Yao, J., Zhu, L., Pu, Y., Chi, C., Sun, H., Wang, B. et al., 2025. Pixel-perfect depth with semantics-prompted diffusion transformers. *arXiv preprint arXiv:2510.07316*.
- Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything v2. *Advances in Neural Information Processing Systems*, 37, 21875–21911.
- Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. Mvsnet: Depth inference for unstructured multi-view stereo. *Proceedings of the European conference on computer vision (ECCV)*, 767–783.