ARAI-MVSNet: A multi-view stereo depth estimation network with adaptive depth range and depth interval

Song Zhang [a,b,c], Wenjia Xu [d], Zhiwei Wei [a,b,*], Lili Zhang [a,b], Yang Wang [a,b], Junyi Liu [a,b]

[a] *Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China*
[b] *Key Laboratory of Network Information System Technology(NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China*
[c] *School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China*
[d] *State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China*

ARTICLE INFO

ABSTRACT

Multi-View Stereo (MVS) is a fundamental problem in geometric computer vision which aims to reconstruct a scene using multi-view images with known camera parameters. However, the mainstream approaches represent the scene with a fixed all-pixel depth range and equal depth interval partition, which will result in inadequate utilization of depth planes and imprecise depth estimation. In this paper, we present a novel multi-stage coarse-to-fine framework to achieve adaptive all-pixel depth range and depth interval. We predict a coarse depth map in the first stage, then an Adaptive Depth Range Prediction module is proposed in the second stage to zoom in the scene by leveraging the reference image and the obtained depth map in the first stage and predict a more accurate all-pixel depth range for the following stages. In the third and fourth stages, we propose an Adaptive Depth Interval Adjustment module to achieve adaptive variable interval partition for pixel-wise depth range. The depth interval distribution in this module is normalized by Z-score, which can allocate dense depth hypothesis planes around the potential ground truth depth value and vice versa to achieve more accurate depth estimation. Extensive experiments on four widely used benchmark datasets (DTU, TnT, BlendedMVS, ETH 3D) demonstrate that our model achieves state-of-the-art performance and yields competitive generalization ability. Particularly, our method achieves the highest Acc and Overall on the DTU dataset, while attaining the highest Recall and F_1-score on the Tanks and Temples intermediate and advanced dataset. Moreover, our method also achieves the lowest e_1 and e_3 on the BlendedMVS dataset and the highest Acc and F_1-score on the ETH 3D dataset, surpassing all listed methods. Project website: https://github.com/zs670980918/ARAI-MVSNet

1. Introduction

Multi-View Stereo (MVS) is a fundamental problem in computer vision that aims to reconstruct a 3D scene from a collection of multiple images captured from different viewpoints. It addresses the challenge of inferring the 3D geometry of a scene by leveraging the visual information contained in these images. The MVS problem has garnered significant attention due to its wide range of applications, including 3D reconstruction [1], robot navigation [2,3] and so on. To tackle the MVS problem, various techniques have been developed over the years. Early traditional MVS methods [4] leveraging multi-view consistency measures have achieved considerable performance, but they still demonstrate limited representation capability in low-texture regions, reflections, etc [5]. Recent advancements in deep learning have revolutionized the field of MVS. MVSNet [5] that utilizes 2D CNN to extract features and uses 3D CNN to regularize the cost volume yields significant improvements on many MVS benchmarks. However,

the pipeline still suffers from low efficiency and high memory consumption. Following works which aim to balance the effectiveness and the efficiency are mainly categorized as recurrent methods [6,7] and multi-stage methods [8,9]. The recurrent methods utilize the GRU or convolutional LSTM to regularize the cost volume sequentially. While the multi-stage methods propose a novel cascade formulation using multi-scale cost volumes. Following a similar cascade strategy, we have designed our framework to achieve an effective and efficient MVSNet.

The basic idea of our framework is that the surfaces of the objects are usually heterogeneous and their sizes are diverse. However, the above methods all represent the object with a fixed all-pixel depth range and equal depth interval partition and will result in two problems. First, since the objects may have different sizes, the fixed all-pixel depth range may not be a suitable fit for the range of each object, as depicted in Fig. 1. As a result, if the fixed all-pixel depth range is larger than the actual depth range of the object, certain depth hypothesis
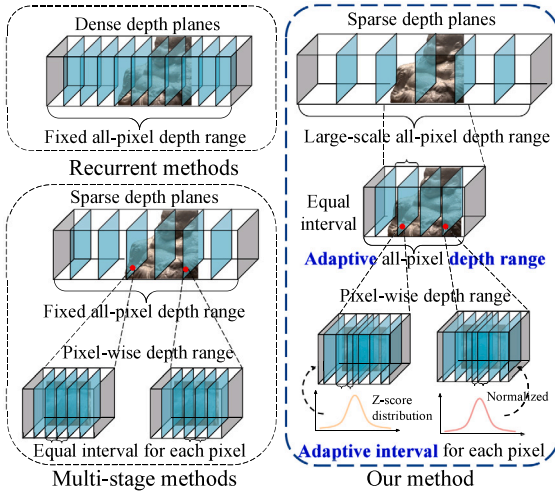

Fig. 1. Comparisons of the recurrent methods, multi-stage methods and our proposed method. We adopt an all-pixel adaptive depth range and adaptive variable interval partition for pixel-wise depth range.

planes may become redundant (the higher left part of Fig. 1), whereas objects that fall beyond the fixed depth range cannot be reconstructed and may lead to a subpar reconstruction quality. Second, since the object surface is usually heterogeneous, employing the equal depth interval partition strategy can result in disparities between predicted depth values and the actual ground truth (GT) depth values (the lower left part of Fig. 1). Insufficient depth hypothesis planes in close proximity to the GT depth value can hinder the ability to accurately estimate depth values, ultimately leading to the imprecise reconstruction of specific details and negatively impacting the overall quality of the reconstruction.

To tackle the aforementioned problems, we propose a novel multi-stage coarse-to-fine framework named ARAI-MVSNet, which mainly consists of two novel modules. As depicted in Fig. 1, instead of using a common fixed all-pixel depth range, we propose an innovative Adaptive Depth Range Prediction (ADRP) module, which enables a more precise zoom-in of the scene. The ADRP module predicts a more accurate all-pixel depth range for each object from a large-scale depth range by leveraging the reference image and the depth map of the former stage. This mechanism effectively utilizes the depth hypothesis planes, leading to the attainment of superior reconstruction outcomes, as evidenced by its high quality. In line with previous studies such as UCS-Net [10] and CFNet [11], we establish the pixel-wise depth range from the all-pixel depth range. However, in contrast to the equal depth interval partition employed in UCS-Net and CFNet, we introduce an Adaptive Depth Interval Adjustment (ADIA) module to reallocate more depth hypothesis planes close to the potential GT depth value. This mechanism utilizes the Z-score formulation [12] to calculate the offset for each depth plane, taking advantage of the depth map from the previous stage to achieve adaptive depth interval partition for the pixel-wise depth range. Additionally, it is also non-parametric to achieve adaptive depth interval for pixel-wise depth range, thus can significantly improve the efficiency of our module with performance maintained. Furthermore, to extract robust image features, we also design an Atrous Spatial Pyramid Feature Extraction Network (ASPFNet), where the context-aware features are extracted and fused with larger receptive fields.

We evaluate our method on four different competitive benchmarks to demonstrate the SOTA performance and generalization ability. The results demonstrate that our model achieves the state-of-the-art Acc score and Overall score compared to other pioneer works on DTU dataset, the highest Recall score and F_1-score score on Tanks and Temples intermediate and advanced dataset respectively. Moreover,

our method also achieve the lowest e_1 score and e_3 score on the BlendedMVS dataset and the highest Acc score and F_1-score on the ETH 3D dataset, surpassing all listed methods.

In summary, our main contributions are four folds: (1) We present a novel multi-stage coarse-to-fine framework to achieve more accurate depth estimation. (2) We propose an ADRP module to predict an adaptive all-pixel depth range for more reliable reconstruction. (3) We propose an ADIA module to achieve adaptive variable depth interval partition to estimate more accurate depth values. (4) We evaluate our method on four benchmarks and significantly advance the state-of-the-art methods in evaluation metrics.

2. Related work

2.1. Traditional MVS methods

The core idea of MVS is to estimate the depth or disparity map for each pixel in the images, which represents the corresponding 3D point's distance from the camera(s). This estimation is achieved by analyzing the spatial and photometric consistencies across the multiple views [13]. The key assumption is that the same 3D point in the scene should project to similar positions in different images, and its appearance should exhibit consistent color or intensity values. By exploiting these correspondences, the MVS algorithms can triangulate the 3D positions of the scene points and reconstruct a dense representation of the underlying geometry [13].

The taxonomic classification of these traditional MVS algorithms can be divided into four main categories: volumetric-based [14], mesh-based [15], point cloud-based [16], and depth map-based [17]. Volumetric, mesh, and point cloud representations directly visualize the scene, while depth map-based methods estimate depth maps of each reference image and fuse them into other representations. This decouples the complex 3D reconstruction problem into a 2D depth map estimation problem, making depth map-based methods more flexible and robust in MVS [18,19]. Despite the fact that traditional MVS methods can obtain impressive results, handcrafted-features based MVS methods still face significant challenges for regions with weak textures and non-Lambertian surfaces [20]. Therefore, our method mainly focuses on further improving the performance of depth estimation by developing a deep learning-based MVS pipeline.

2.2. Deep learning-based MVS methods

Recently, to overcome the blemish of traditional MVS methods, many deep learning-based methods [5,21,22] have been proposed and achieve impressive performances. MVSNet [5] implements a differentiable homography warping to transform the traditional MVS domain into the deep-learning based domain. This approach involves matching the feature points between different views to achieve spatial and photometric consistencies (it constructs a cost volume) across multiple views. Subsequently, 3D CNNs are employed to regularize the cost volume. However, the 3D CNNs greatly increases memory consumption. R-MVSNet [6] and D^2HC-RMVSNet [7] leverage convolutional GRUs and LSTMs to sequentially regularize the cost volume to avoid using memory-intensive 3D CNNs. However, these methods require a large number of depth hypothesis planes to obtain better performance. And it is difficult to construct a high precision cost volume based on the high-resolution image limited by the memory bottleneck. Therefore, multi-stage methods [8,9] have been proposed to obtain high-quality depth maps without increasing memory. UCS-Net [10] uses adaptive thin volumes based on uncertainties of pixel-wise depth predictions. ADR-MVSNet [21] introduces an Adaptive Depth Reduction module that utilizes the probability distribution of pixels in the depth direction to dynamically learn distinct depth ranges for each pixel. Prior-Net [22] uses a three-stage cascade strategy in its fast version to efficiently compute a low-resolution depth map. Additionally, a Refine-Net is
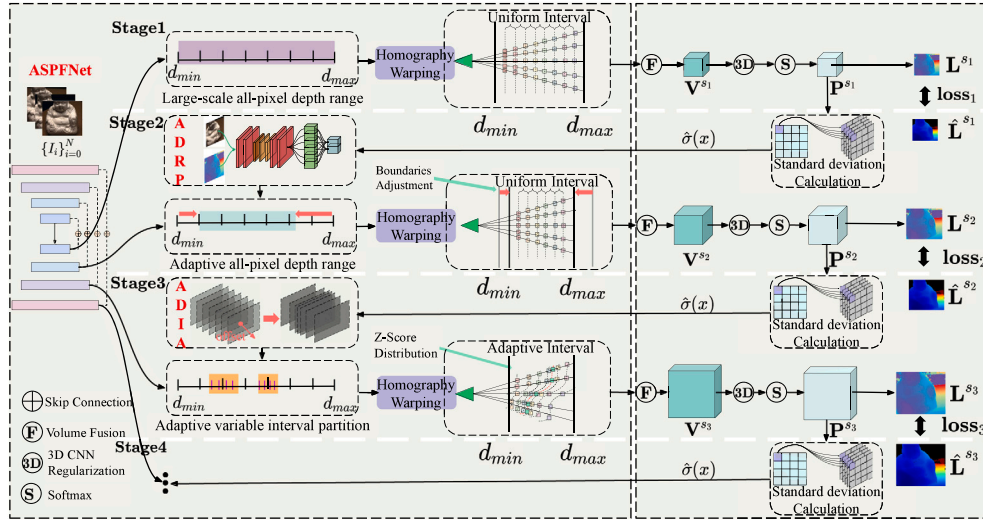
Fig. 2. Illustration of overall ARAI-MVSNet. This is a typical multi-stage coarse-to-fine framework. The left part is our novel modules, and the right part is inherited from existing methods. The three dots represent the stage 4 is a repeat of the stage 3 process.

proposed to upsample the depth and confidence maps to the original high-resolution, further enhancing computational efficiency. Nevertheless, learning-based MVS methods still face challenging issues, such as resource redundancy and limited capability in low-texture regions.

3. Methodology

3.1. Problem formulation

In the deep learning-based MVS task, the goal is to use an end-to-end trainable model to infer a depth map $\mathbf{L}$ from $N-1$ adjacent views with their corresponding camera poses. Assume the reference image $\mathbf{I}_1$ and source images $\{\mathbf{I}_i\}_{i=2}^{N}$, where the features $\{\mathbf{F}_i\}_{i=1}^{N}$ are extracted from them. Then the differentiable homography warping based on the depth hypothesis planes $\mathbf{D}$ and the extracted $\{\mathbf{F}_i\}_{i=1}^{N}$ is used to construct a cost volume $\mathbf{V}$. After regularization process, the regularized cost volume $\mathbf{V}_{reg}$ regresses a probability volume $\mathbf{P}$ by softmax operation. Finally, the global depth map $\mathbf{L}$ is calculated by the probability volume $\mathbf{P}$ and depth hypothesis planes $\mathbf{D} = [\mathbf{d}_{min}, \dots, \mathbf{d}_{max}]$. And $\mathbf{d}$ represents a depth value among the depth hypothesis planes.

3.2. ARAI-MVSNet

The overall framework of ARAI-MVSNet is shown in Fig. 2. It employs a four-stage coarse-to-fine approach for depth map inference. Prior to entering the pipeline, a feature extraction network called ASPFNet is used to extract multi-scale context features. **Stage 1:** we use the main part of MVSNet based on a large-scale all-pixel depth range to obtain a coarse depth map. **Stage 2:** we propose an Adaptive Depth Range Prediction (ADRP) module by leveraging the reference image and the obtained coarse depth map in stage 1 to predict an adaptive all-pixel depth range, then the standard flow is executed as above based on the new all-pixel depth range. **Stage 3:** we propose an Adaptive Depth Interval Adjustment (ADIA) module to adaptively adjust the depth interval for pixel-wise depth range based on Z-score distribution. Then the standard flow is executed as above based on the new depth interval partition. **Stage 4:** Stage 4 is similar to Stage 3, but with a greater number of depth hypothesis planes for improved performance.

3.3. Atrous spatial pyramid feature extraction network (ASPFNet)

Conventional CNNs face challenges in effectively handling reflective surfaces, low-textured regions, and texture-less regions when operating on standard 2D grids with fixed receptive fields [20]. Therefore, to address this challenge and extract more comprehensive image features, we have designed an Atrous Spatial Pyramid Feature Extraction Network (ASPFNet) that combines dilated convolution and FPN to enlarge the receptive field and obtain better multi-scale features. The ASPFNet is an encoder–decoder architecture (as shown in Fig. 2). The ASPFNet is composed of four Encoder Downsample Blocks (EDB), four Decoder UpSample Blocks (DUB) and skip connections. EDB shrinks the feature map and enlarges the receptive field with dilated convolution. DUB enlarges the feature map and fuses high-dimensional features from skip connections. This bottleneck encoder–decoder architecture can integrate multi-level features by simultaneously fusing low- and high-level information to achieve high-quality reconstruction. More details of ASPFNet are in the supplementary material.

3.4. Adaptive depth range prediction (ADRP)

To achieve the image-level uniform adjustment, we utilize a large-scale all-pixel depth range to obtain a coarse depth map in stage 1. Since the large-scale all-pixel depth map may not be accurate enough to support subsequent stages, we propose an Adaptive Depth Range Prediction (ADRP) module in stage 2 that calculates an adaptive all-pixel depth range by leveraging the information from the coarse depth map of stage 1 and the reference image.

In stage 1, we establish a large-scale all-pixel depth range to derive a preliminary coarse depth map denoted as $\mathbf{L}^{s_1}$. The maximum and minimum depth values of $\mathbf{L}^{s_1}$, represented by $\mathbf{L}^{s_1}(x_{min})$ and $\mathbf{L}^{s_1}(x_{max})$, respectively, are then utilized as the initial depth boundaries in stage 2. Where the x represents a pixel in the depth map, x_{max} and x_{min} denote the position of the pixel of the maximum and minimum depth values in the depth map, respectively. Using the reference image $\mathbf{I}_{ref}$ and coarse depth map $\mathbf{L}^{s_1}$ as inputs, we predict an updated and more precise all-pixel depth range by designing a Scalar Calculation Network (SCN) (illustrated in Fig. 3) to calculate α_{dr} and β_{dr} for adjusting the depth boundaries:

$$\mathbf{d}_{min}^{s_2} = \mathbf{L}^{s_1}(x_{min}) + \alpha_{dr} \times \hat{\sigma}(x_{min})$$
$$\mathbf{d}_{max}^{s_2} = \mathbf{L}^{s_1}(x_{max}) + \beta_{dr} \times \hat{\sigma}(x_{max})$$

(1)
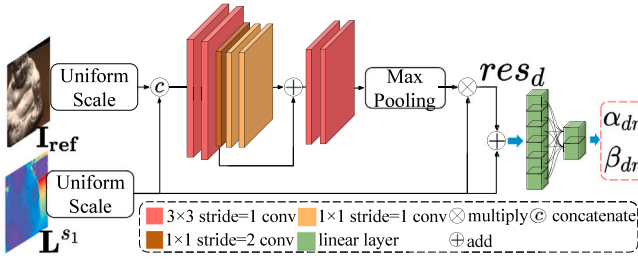
Fig. 3. Illustration of Scalar Calculation Network.

Fig. 4. The scheme of Adaptive Depth Interval Adjustment Module.

where $\hat{\sigma}(x_{min})$ and $\hat{\sigma}(x_{max})$, working as the adjustment factors, are the standard deviations corresponding to positions. As the loss decreases, our depth estimation becomes more accurate, which reduces $\hat{\sigma}(x_{min})$ over time. Thus, by adjusting the coarse depth range while maintaining the loss constraint, we can predict a more precise adaptive all-pixel depth range.

In particular, we consider the depth value as the mean value of previous stage, and $\hat{\sigma}(x_{min})$ and $\hat{\sigma}(x_{max})$ are calculated as Eq. (2).

$$\hat{\sigma}(x_i) = \sqrt{\sum_j^D \mathbf{P}_j^{s_1}(x_i) \cdot \left(\mathbf{d}_j^{s_1}(x_i) - \mathbf{L}^{s_1}(x_i)\right)^2} \tag{2}$$

with $x_i \in [x_{min}, x_{max}]$, where $\mathbf{P}_j^{s_1}$ expresses pixel-wise depth probability distributions of stage 1, $\mathbf{d}_j^{s_1}$ represents the depth hypotheses plane of stage 1, and $\mathbf{P}_j^{s_1}(x_i)$ represents how probable the depth at pixel x_i is $\mathbf{d}_j^{s_1}(x_i)$.

We concatenate the coarse depth map and reference image, and leverage stacked convolutional layers $\mathbf{w}_1$ to extract depth residual information res_d. We add res_d to the depth map, and regress the final scalars α_{dr}, β_{dr} with a linear layer $\mathbf{w}_2$:

$$\begin{aligned} res_d &= maxp(\mathbf{w}_1([\mathbf{L}^{s1}, \mathbf{I_{ref}}]) \odot \mathbf{L}^{s1}), \\ \alpha_{dr}, \beta_{dr} &= \mathbf{w}_2(res_d + \mathbf{L}^{s1}), \end{aligned} \tag{3}$$

After obtaining the adaptive all-pixel depth range, we partition the all-pixel depth range by equal interval partition to obtain depth hypothesis planes $\mathbf{D}^{s2} = [\mathbf{d}_{min}^{s2}, \dots, \mathbf{d}_{max}^{s2}]$ to construct the cost volume in stage 2. Based on the cost volume, then a cost volume regularization network (similar to the 3D UNet used in MVSNet [5]) is employed to regress a depth map $\mathbf{L}^{s2}$ in stage 2.

3.5. Adaptive depth interval adjustment (ADIA)

Instead of employing equal interval partition for pixel-wise depth estimation, we propose an adaptive variable interval partition strategy to allocate more dense depth hypothesis planes near the potential ground truth depth value, and allocate relatively sparse depth hypothesis planes away from it. Z-score is a measure of the relative position, which can measure the distance from the values to the mean (namely the GT depth value). Thus, we follow the Z-score distribution [12] to allocate the position of depth hypothesis planes to achieve the above strategy. To accomplish this, we propose an Adaptive Depth Interval Adjustment (ADIA) module (as illustrated in Fig. 4). The computation procedure is delineated in Eq. (4), in which we calculate different weights for equal intervals to obtain variable intervals and add them to the corresponding basic depth hypothesis planes to accomplish adaptive depth interval adjustment.

$$\mathbf{d}_i^{s_3}(x) = \mathbf{d}_i^{s_3}(x) + \hat{\mathbf{d}}_{inter}^{s_3}(x) \times \mathbf{offset}_i(x) \tag{4}$$

where $\mathbf{d}_i^{s_3}(x)$ represents ith plane of depth hypothesis planes of pixel x, $\hat{\mathbf{d}}_{inter}^{s_3}(x)$ is the pixel-wise equal interval, $\mathbf{offset}_i(x)$ is the weight for the equal interval. In Eq. (4), we allocate different offsets to depth intervals to achieve interval scaling based on the Z-score [12] formulation.
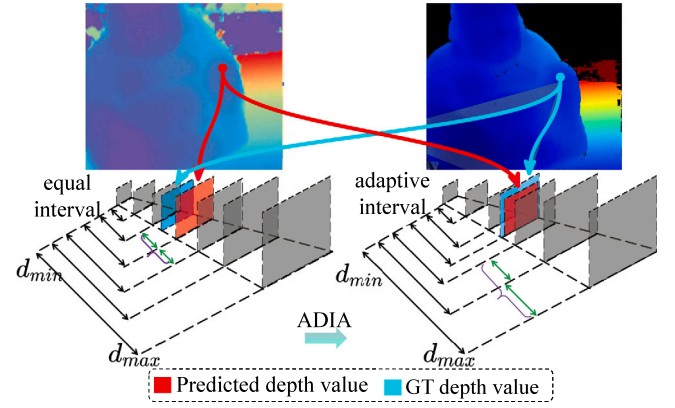
The calculation process of $\hat{\mathbf{d}}_{inter}^{s_3}(x)$ and $\mathbf{offset}_i(x)$ is outlined as follows. Specifically, we first utilize previous stage depth map $\mathbf{L}^{s2}$ and probability volume $\mathbf{P}^{s2}$ to calculate the pixel-wise standard deviation $\hat{\sigma}(x)$, which is defined as Eq. (5).

$$\hat{\sigma}(x) = \sqrt{\sum_j^D \mathbf{P}_j^{s_2}(x) \cdot \left(\mathbf{d}_j^{s_2}(x) - \mathbf{L}^{s_2}(x)\right)^2} \tag{5}$$

Then we leverage above results to calculate the pixel-wise depth range for current stage, the upper and lower boundaries of the pixel-wise depth range, i.e., pixel-wise depth min $\mathbf{d}_{min}^{s3}(x)$ and pixel-wise depth max $\mathbf{d}_{max}^{s3}(x)$ are defined as Eq. (6).

$$\begin{aligned} \mathbf{d}_{min}^{s_3}(x) &= \mathbf{L}^{s_2}(x) - \hat{\sigma}(x), \mathbf{d}_{max}^{s_3}(x) = \mathbf{L}^{s_2}(x) + \hat{\sigma}(x) \\ \mathbf{D}^{s3}(x) &= [\mathbf{d}_{min}^{s_3}(x), \dots, \mathbf{d}_i^{s_3}(x), \dots, \mathbf{d}_{max}^{s_3}(x)] \end{aligned} \tag{6}$$

where $\mathbf{D}^{s3}(x)$ represents the pixel-wise depth hypothesis planes of stage 3. And we partition above pixel-wise depth range to obtain the pixel-wise equal interval $\hat{\mathbf{d}}_{inter}^{s_3}(x)$ as our fixed step. The standard equal interval partition is defined as Eq. (7).

$$\hat{\mathbf{d}}_{inter}^{s_3}(x) = \frac{\mathbf{d}_{max}^{s_3}(x) - \mathbf{d}_{min}^{s_3}(x)}{\mathbf{D}_{num}} \tag{7}$$

where $\mathbf{D}_{num}$ represents the number of depth hypothesis planes of stage 3. Inspired by Z-score [12], we utilize the depth value of the previous stage (considered as mean value) and standard deviation to calculate the offset for each plane, and use softmax to achieve normalization. The calculation process is formulated by Eq. (8).

$$\mathbf{offset}_i(x) = softmax\left(\frac{\mathbf{d}_i^{s_3}(x) - \mathbf{L}^{s_2}(x)}{\hat{\sigma}(x)}\right) \tag{8}$$

where $\mathbf{d}_i^{s_3}(x)$ represents the ith depth value correspondent to pixel x among the depth hypothesis planes of stage 3, and $\mathbf{L}^{s_2}(x)$ represents the depth value of pixel x in predicted depth map of previous stage. Finally, we use the offset to achieve adaptive variable interval partition, as defined by Eq. (4).

3.6. Training loss

Following the previous methods [5], we adopt same mean absolute difference loss as our loss:

$$Loss = \sum_{i=1}^4 \lambda_i \sum_{x \in \mathbf{x}_{valid}} \|\hat{\mathbf{L}}^{s_i}(x) - \mathbf{L}^{s_i}(x)\|_1, \tag{9}$$

where $\hat{\mathbf{L}}^{s_i}$ denotes the GT depth map of each stage, $\mathbf{L}^{s_i}$ denotes the predicted depth map of each stage. We set λ_i to be 0.5, 1.0, 1.5, 2.0 for each stage.
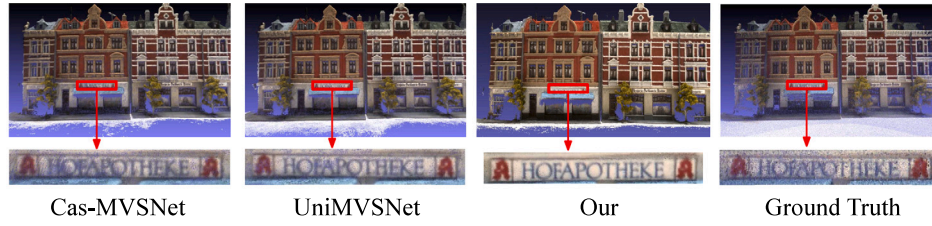
| Cas-MVSNet | UniMVSNet | Our | Ground Truth |

Fig. 5. Quantitative results of scan 15 on DTU. The top row exhibits the point clouds obtained by different methods. The bottom row exhibits the detailed comparison.

4. Experiments

4.1. Dataset

We evaluate our method on four datasets widely used in MVS. These include: (1) **DTU** [23], captured in a laboratory setting, it consists of 124 scenes and 7 lighting conditions. (2) **Tanks and Temples** [24], captured from real outdoor sensors, with more complex and realistic scenes; (3) **BlendedMVS** [25], with over 17k indoor and outdoor images of 113 scenes split into training and testing sets; and (4) **ETH 3D** [26], with high-resolution calibrated images of scenes containing significant viewpoint variations.

4.2. Implement details

Training: The proposed ARAI-MVSNet was trained on DTU dataset and finetuned on BlendedMVS dataset. (1) During training, input image resolution was set to 640×512 and $N = 3$ training views were used. Hypothetical depth planes of four stages were set to $[16, 48, 16, 8]$. The model was optimized for 16 epochs using RMSProp optimizer with a learning rate of 0.0006 and decay weight of 0.001. The batch size was 16 and 8 NVIDIA GTX 2080ti GPUs were used for training. Standard metrics provided by the official evaluation protocol [23] were used for performance evaluation. (2) For finetuning, the best result of DTU was used for 10 epochs with $N = 7$ images of original size of 768×576. The learning rate was set to 0.0005 and the hypothetical depth plane was set to $[16, 64, 32, 8]$. **Testing**: (1) For testing on DTU dataset, our best result was used with $N = 7$ adjacent images, image resolution of 960×1280, and hypothetical depth planes of $[16, 64, 16, 8]$. (2) To verify the generalization ability on Tanks and Temples benchmark, the finetuned model was used with input image sizes of 1024×1920 or 1024×2048, $N = 11$ input images, and hypothetical depth planes of $[16, 128, 32, 16]$.

4.3. Performance on close-range experimental scene benchmark

Close-range scenarios are characterized by a short distance between the camera and the scene being captured. Given that the DTU dataset contains small scenes and close camera-to-scene distances, we leveraged this dataset to evaluate the effectiveness of our model on a close-range experimental scene benchmark. Table 1 shows the quantitative results on DTU dataset.

From Table 1, we have the following observations: our ARAI-MVSNet achieves the SOTA Acc score. Specifically, our ARAI-MVSNet improves the Acc from 0.325 of CasMVSNet [8] to 0.292, 0.354 of ADR-MVSNet [21] to 0.292, 0.351 of Prior-Net [22] to 0.292 and 0.352 of UniMVSNet [30] to 0.292 (1st), meanwhile higher than previous deep learning-based SOTA CVP-MVSNet [9] (0.292 vs. 0.296). The reason for this is that the DTU dataset inherently falls under the close-range category and consists of highly heterogeneous and fine-grained object surfaces. Our ADIA module addresses this challenge by employing a Z-score mechanism, allowing it to allocate more depth hypothesis planes in the finer regions for accurate depth estimation. The qualitative results of our method compared to above methods in Fig. 5 also

Table 1
Quantitative results of different methods on DTU's evaluation set.

Methods	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
MVSNet [5]	0.396	0.527	0.462
R-MVSNet [6]	0.383	0.452	0.417
CasMVSNet [8]	0.325	0.385	0.355
CVP-MVSNet [9]	0.296	0.406	0.351
UCS-Net [10]	0.338	0.349	0.344
D^2HC-RMVSNet [7]	0.395	0.378	0.386
AA-RMVSNet [20]	0.376	0.339	0.357
PatchmatchNet [27]	0.427	0.277	0.352
EPP-MVSNet [28]	0.413	0.296	0.355
ADR-MVSNet [21]	0.354	0.317	0.335
IterMVS [29]	0.373	0.354	0.363
UniMVSNet [30]	0.352	0.278	0.315
NP-CVP-MVSNet [31]	0.356	**0.275**	0.315
Prior-Net [22]	0.351	0.287	0.319
Our	**0.292**	0.334	**0.313**

convince the above observations. Notably, our model produces more precise point cloud reconstructions on object surfaces, as indicated by the outlying rectangle in Fig. 5. Moreover, our model also ranks first in Overall performance (0.313, 1st) among state-of-the-art methods, e.g. UniMVSNet [30], NP-CVP-MVSNet [31], and Prior-Net [22].

4.4. Performance on large-scale practical scenes benchmark

Large-scale represents that the shooting distance between the camera and the scene is relatively far. Therefore, we adopted the benchmark composed of camera-captured or synthetic views in wild environments to evaluate the performance of our model on the large-scale practical scenes benchmarks, e.g. Tanks and Temples benchmark, BlendedMVS benchmark, etc.

Evaluation on Tanks and Temples benchmark: We executed a test on Tanks and Temples to evaluate the effectiveness of our model in large-scale dataset. We adopted the same dynamic geometric consistency checking strategies as D^2HC-RMVSNet [7], and chose an appropriate mask threshold for each scene. The quantitative results in Table 2 indicates that our ARAI-MVSNet achieves competitive performance among all listed MVS methods and obtains SOTA results in some scenes. Our findings are as follows: our method achieves the best Recall score (79.90) and competitive results in terms of F_1-score and Precision. Notably, our ARAI-MVSNet improves the Recall score from 75.69 in AA-RMVSNet [20] to 79.90, from 63.94 in ADR-MVSNet [21] to 79.90, from 76.20 in Prior-Net [22] to 79.90, and from 73.82 in UniMVSNet [30] to 79.90 (1st), while maintaining comparable F_1-score and precision with all other MVS methods listed. Moreover, our method also obtain the highest F_1-score in Family (66.83) and Horse (54.15) compared with all listed methods. The Tanks and Temples benchmark dataset comprises large-scale scenes, wherein the objects of interest occupy only a fraction of the entire scene. Our proposed Adaptive Depth Range Prediction (ADRP) module, equipped with an adaptive boundaries adjustment mechanism, enables accurate prediction of the all-pixel depth range by zooming in on the scene, resulting in more comprehensive reconstruction results. We conclude that the improvement in Recall is due to the mechanism of the ADRP.

Table 2
Generalization results on the TnT. The left part lists our F_1-scores on each scene. The right part lists the mean values of F_1-score, Precision and Recall.

Methods	Fam. ↑	Fra. ↑	Hor. ↑	Lig. ↑	M60 ↑	Pan. ↑	Pla. ↑	Tra. ↑	F_1-score ↑	Precision ↑	Recall ↑
MVSNet [5]	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69	43.48	40.23	49.70
R-MVSNet [6]	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38	48.40	43.74	57.60
CasMVSNet [8]	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51	56.84	47.62	74.01
CVP-MVSNet [9]	76.5	47.74	36.34	55.12	57.28	54.28	57.43	47.54	54.03	51.41	60.19
UCS-Net [10]	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89	54.83	46.66	70.34
D^2HC-RMVSNet [7]	74.69	56.04	49.42	60.08	59.81	59.61	60.04	53.92	59.20	49.88	74.08
AA-RMVSNet [20]	77.77	59.53	51.53	**64.02**	64.05	59.47	60.85	55.50	61.51	52.68	75.69
EPP-MVSNet [28]	77.86	60.54	52.96	62.33	61.69	60.34	**62.44**	55.30	61.68	53.09	75.58
ADR-MVSNet [21]	72.61	52.81	37.43	53.46	51.88	48.71	59.11	43.76	52.47	46.62	63.94
IterMVS [29]	76.12	55.80	50.53	56.05	57.68	52.62	55.70	50.99	56.94	47.53	74.69
UniMVSNet [30]	**81.20**	66.43	53.11	63.46	**66.09**	**64.84**	62.23	**57.53**	**64.36**	**57.54**	73.82
Prior-Net [22]	79.02	60.93	51.65	60.52	61.78	56.19	61.37	53.59	60.63	50.80	76.20
ARAI-MVSNet	79.48	**66.83**	**54.15**	59.56	58.58	57.38	56.51	56.27	61.09	50.42	**79.90**

Table 3
The F_1 score on Tanks and Temple advanced.

Methods	Mean	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
CasMVSNet [8]	31.12	19.81	38.46	29.10	43.87	27.36	28.11
AA-RMVSNet [20]	33.53	20.96	40.15	32.05	46.01	29.28	32.71
IDCF [32]	32.28	23.66	33.01	36.05	42.10	23.95	34.92
EPP-MVSNet [28]	35.72	21.28	39.74	35.34	**49.21**	30.00	38.75
ADR-MVSNet [21]	26.75	20.06	26.60	27.92	35.28	26.85	23.77
TransMVSNet [33]	37.00	24.84	**44.59**	34.77	46.49	34.69	36.62
Prior-Net [22]	34.61	**26.99**	40.46	30.76	47.81	28.96	32.71
Our	**38.68**	26.13	43.01	**38.63**	48.88	**35.39**	**40.01**

We also executed the evaluation on TnT advanced dataset (Table 3). Our findings are as follows: our model achieves the best Mean F_1-score than all listed methods. In particular, our model improves the F_1-score from 33.53 in AA-RMVSNet [20] to 38.68, from 32.28 in IDCF [32] to 38.68, from 26.75 in ADR-MVSNet [21] to 38.68, from 34.61 in Prior-Net [22] to 38.68, and from 37.00 in TransMVSNet [33] to 38.68 (1st). The F_1-score is calculated from both Precision and Recall. Therefore, achieving SOTA F_1-score means that our model performs exceptionally well in both Precision and Recall. The outstanding results are mainly attributed to the superior performance of our proposed ADIA module in Precision and the excellent performance of our ADRP module in Recall. Besides, our method can obtain the highest results on several scenes, i.e. Courtroom, Palace and Temple.

Evaluation on BlendedMVS benchmark: As shown in Table 4 left, we tested our model on BlendedMVS [25] validation set to further demonstrate the generalizability and flexibility of ARAI-MVSNet. We can observe that our method achieves the highest scores in e_1 and e_3. Specially, we improve the e_1 from 12.66 in EPP-MSVNet [28] to 7.91 and from 9.35 in UniMVSNet [30] to 7.91. Moreover, we also improve the e_3 from 6.20 in EPP-MSVNet [28] to 2.95 and from 3.25 in UniMVSNet [30] to 2.95. e_1 and e_3 represent the proportion in % of pixels with depth error larger than 1 and larger than 3, the SOTA performances on e_1 and e_3 confirm the advantage of our ADIA module as ADIA can estimate more accurate depth values by reallocating the depth planes. Our ADRP module pays more attention to the depth range of the foreground but may ignore the depth range of the background, thus, the calculation of EPE may be affected by outliers. Therefore, our EPE is slightly lower than UniMVSNet [30] (0.67 vs. 0.62), but higher than all other models.

Evaluation on ETH 3D benchmark: We conducted further experiments on the large-scale scene dataset (ETH 3D [26]) to validate the performance of ARAI-MVSNet in the realistic wild environment. We can observe that our method obtain the highest Acc and F1-score on ETH 3D (as shown in Table 4, right)). In particular, we have the following observations: our ARAI-RMVSNet improves the Acc from 74.48 of PatchMatch-RL [27] to 92.12 and 84.73 of Iter-MVS-RL [30] to 92.12. And our ARAI-RMVSNet also improves the F_1-score from 72.38 of PatchMatch-RL [27] to 84.90 and 80.06 of Iter-MVS-RL [30] to

84.90. Our model achieves SOTA performance on Acc when benefitting from the accurate depth planes reallocated by ADIA. However, we have noticed that different lighting conditions in the ETH 3D dataset, unlike DTU and BlendedMVS datasets used for training, may cause our ADRP module to estimate an incorrect all-pixel depth range due to dark lighting. This incorrect estimation can adversely impact the reconstruction results on Comp. Moreover, the SOTA Acc and competitive Comp of our model enable it to achieve SOTA F_1-score. These indicate that our method significantly improves overall reconstruction quality and demonstrates competitive generalization ability compared to most deep learning-based MVS methods.

4.5. Comparison of the models efficiency

In this section, we firstly compared the inference time with several multi-stage MVS methods on the DTU dataset. The results in Table 5 indicate that our method produces accurate point clouds with lower runtime than previous methods, with reductions of 0.9 s, 0.16 s, and 0.27 s compared to CVP-MVSNet [9], UCS-Net [10], and UniMVSNet [30], respectively. Our model also has competitive memory usage (5386 MB) compared to mainstream methods [8–10,30]. Although our four-stage framework has slightly higher memory usage than UCS-Net, it performs better in other efficiency metrics. Compared to UniMVS-Net, which has the best Comp score of 0.2912, our model achieves better scores in terms of Acc and Overall performance and has a faster inference time. The non-parametric mechanism in our ADIA module significantly improves the inference time. Furthermore, we additionally compared our method with some superior Transformer-based methods (e.g. MVSFormer [34], WT-MVSNet [35]). The quantitative results in Table 6 reveal that our method achieves better or comparable results (highest Recall: 79.90) with lower memory and time consumption than Transformer-based methods, with reductions of 61% and 18% in memory and 8% and 44% in time compared to MVSFormer [34] and WT-MVSNet [35]. We attribute our model's efficiency superiority to two main reasons. Firstly, methods like MVS-Former, WT-MVSNet, etc., being Transformer-based, inherently demand more computation and memory resources compared to ARAI-MVSNet, which adopts a simpler and more lightweight CNN-based framework. Secondly, the ADIA module proposed in our work utilizes a

Table 4
Left part: The results on BlendedMVS. EPE stands for the endpoint error, which is the average ζ-1 distance between the prediction and the ground truth depth; e_1 and e_3 represent the proportion in % of pixels with depth error larger than 1 and larger than 3. **Right part**: The evaluation results on ETH 3D are obtained from Iter-MVS.

Methods	e_1 ↓	e_3 ↓	EPE ↓	Methods	Acc ↑	Comp ↑	F_1-score ↑
MVSNet [5]	21.98	8.32	1.49	Gipuma [4]	86.47	24.91	45.18
UCS-MVSNet [10]	14.12	7.33	1.32	PatchmatchNet [27]	69.71	77.46	73.12
EPP-MVSNet [28]	12.66	6.20	1.17	Iter-MVS [29]	76.91	72.65	74.29
TransMVSNet [33]	8.32	3.62	0.73	PatchMatch-RL [27]	74.48	72.06	72.38
UniMVSNet [30]	9.35	3.25	**0.62**	Iter-MVS-RL [29]	84.73	76.49	80.06
Our	**7.91**	**2.95**	0.67	Our	**92.12**	78.73	**84.90**

Table 5
Performance comparisons on the DTU val set. "Time" is the inference time.

Methods	Input size	View	Num depth	Acc (mm) ↓	Comp (mm) ↓	Overall (mm) ↓	GPU Mem (MB)	Time (s)
CasMVSNet [8]	1280 × 960	7	[64,16,8]	0.3508	0.4039	0.37735	4939 MB	0.64 s
CVP-MVSNet [9]	1280 × 960	7	–	0.3023	0.4256	0.36395	6327 MB	1.51 s
UCS-Net [10]	1280 × 960	7	[64,16,8]	0.3609	0.3764	0.36865	**4889 MB**	0.77 s
UniMVSNet [30]	1280 × 960	7	[64,16,8]	0.3772	**0.2912**	0.3342	6441 MB	0.88 s
ARAI-MVSNet	1280 × 960	7	[16,64,16,8]	**0.2924**	0.3342	**0.3133**	5386 MB	**0.61 s**

Table 6
Quantitative results on TnT (Left) and efficiency results on DTU (Right). The results of WT-MVSNet are from the official paper since the code is not open source.

Methods	F_1-score ↑	Precision ↑	Recall ↑	Input size	View	Memory (MB) ↓	Time (s) ↓
MVSFormer	**66.41**	**60.77**	74.33	864 × 1152	7	11 662 MB	0.614 s
WT-MVSNet	65.34	56.09	79.04	864 × 152	7	5549 MB	0.978 s
ARAI-MVSNet (Ours)	61.09	50.42	**79.90**	864 × 1152	7	**4562 MB**	**0.543 s**

Table 7
The overlap rate of predicted and fixed all-pixel depth range. F-score is the harmonic mean of AOG and AOS.

Methods	Mean AOG ↑	Mean AOS ↑	F-score ↑
$\mathbf{D}_{128}$	55.56%	57.46%	56.49%
$\mathbf{D}_{256}$	**98.58%**	56.33%	71.69%
$\mathbf{D}_{adrp}$ (Ours)	90.88%	**79.27%**	**84.68%**

Table 8
The total number of pixels with predicted values close to the GT depth value.

Methods	Modules			P_{numD} ↑
	ASPFNet	ADRP	ADIA	
− ADIA	✓	✓		32.97%
+ ADIA	✓	✓	✓	**61.76%**

simple and non-parametric mechanism (Z-score), introducing minimal additional memory and computational overhead while still delivering performance improvements.

4.6. Quantitative analysis of depth estimation results

The accuracy of the depth map directly affects the final reconstructed point cloud quality. Our model has two novel modules, ADRP and ADIA. Specifically, ADRP predicts an adaptive depth range to zoom in a scene, while ADIA achieves adaptive depth interval partition by reallocating the depth hypothesis planes through offset learning, resulting in more accurate depth values. The effectiveness of these modules were also evaluated based on the depth estimate results, as follows:

Comparison with depth range overlap rate: In this section, we compared the overlap rate of different all-pixel depth ranges predicted by ADRP, with depth hypothesis planes set to 128 and 256, with the ground truth all-pixel depth range. The results are shown in Table 7. From the table, we can observe that our predicted adaptive all-pixel depth range can cover objects more comprehensively than the fixed all-pixel depth range while using fewer depth hypothesis planes. Suppose $\mathbf{D}_{adrp}$ is the predicted all-pixel depth range $\mathbf{D}_{adrp}$ which is obtained by ADRP, $\mathbf{D}_{128}(\mathbf{D}_{num} = 128)$ and $\mathbf{D}_{256}(\mathbf{D}_{num} = 256)$ are the two default all-pixel depth range in most models [5,6]. To evaluate whether our ADRP module can obtain a more accurate all-pixel depth range for the scene reconstruction, two metrics to measure the overlap between

the adopted all-pixel depth range and the GT all-pixel depth range $\mathbf{D}_{gt}$ obtained by the GT depth map are defined:

$$\text{AOG} = \frac{|\mathbf{D}_{gt} \cap \mathbf{D}_i|}{|\mathbf{D}_{gt}|}, \quad \text{AOS} = \frac{|\mathbf{D}_{gt} \cap \mathbf{D}_i|}{|\mathbf{D}_i|} \tag{10}$$

where $\mathbf{D}_i \in [\mathbf{D}_{adrp}, \mathbf{D}_{128}, \mathbf{D}_{256}]$, $\cap$ represents the intersection operation and $|\cdot|$ is the length of the depth range. A higher AOG represents a higher overlap rate with the GT all-pixel depth range, while a higher AOS denotes less waste in the all-pixel depth range.

The mean results on the DTU evaluation set are shown in Table 7. The results indicate that by setting a smaller number of depth hypothesis planes $\mathbf{D}_{128}$, the fixed all-pixel depth range (e.g. $[425, 425 + 2.5 \times 128]$) cannot cover enough effective space for the scene (reflected on MAOG = 55.56% and MAOG = 57.46%), resulting in low performance. Meanwhile, even if the number of depth hypothesis planes is set to be a larger value (e.g. $\mathbf{D}_{num} = 256$) to enlarge the all-pixel depth range to cover the GT all-pixel depth range (MAOG = 98.58%), the overhead of the model will increase and lead to the wastage of the depth planes, i.e., only 56.33% of the depth planes are actually useful. On contrary, our model covers 90.88% of the GT scene and with more percentage of depth hypothesis planes being utilized, i.e., MAOS = 79.27%, which proves that our ADRP module can predict an adaptive and accurate all-pixel depth range close to the GT all-pixel depth range.

Comparison with total accurate number of pixels: As we mentioned in aforementioned section, the calculation of pixel-wise depth range in the current phase involves using the variance of the previous phase's depth values, as defined in Eq. (6), after achieving the equal interval partition. However, employing the equal depth interval partition strategy may lead to inadequate depth hypothesis planes due

Table 9

Left part: Ablation results on DTU evaluation set. The left part of table lists the mean values of Acc, Comp and Overall.
Right part: Ablation results on TnT dataset. The right part of table lists the mean values of F_1-score, Precision and Recall.

Methods	Modules			Acc↓	Comp↓	Overall↓	F_1-score ↑	Precision ↑	Recall ↑
	ASPFNet	ADRP	ADIA						
- ADIA	✓	✓		0.3667	0.3693	0.3680	57.33	45.48	77.53
- ADRP	✓		✓	0.3021	0.3982	0.3502	58.16	48.50	72.63
- ASPFNet		✓	✓	0.3143	0.3532	0.3338	–	–	–
ARAI-MVSNet	✓	✓	✓	**0.2924**	**0.3342**	**0.3133**	**61.09**	**50.42**	**79.90**

Table 10

The ablation study to verify the effectiveness and superiority of our ADRP.

Methods	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓	F_1-score ↑	Precision ↑	Recall ↑
ARAI-MVSNet	0.2924	**0.3342**	**0.3133**	**61.09**	**50.42**	**79.90**
replace ADRP	**0.2901**	0.3402	0.3152	58.46	48.11	74.48

to the heterogeneous nature of object surfaces, resulting in imprecise depth estimation outcomes. Therefore, in this section, we conducted a comparative experiment to verify the effectiveness of our ADIA. Based on the pixel-wise depth range, we have examined two scenarios: one where we employed the adaptive variable interval partition and another employed the equal interval partition. Then we calculated the total number of pixels where the predicted depth value was close to the GT depth value to validate whether our ADIA can achieve more accurate depth estimation for each pixel. Specifically, we calculated the distance (D_e) between the predicted depth value and the GT depth value for all pixels in DTU evaluation set, and displayed the percentage of the number of pixels (P_{numD}) which were close to the ground truth (i.e., $D_e \leq 1e-1$) in Table 8. The results demonstrate that applying the adaptive variable interval partition improves the P_{numD} by a large margin, which indicates that our ADIA module has estimated more accurate depth values for each pixel by setting adaptive depth hypothesis planes. In conclusion, achieving adaptive variable interval partition based on the pixel-wise depth range is imperative.

4.7. Ablation experiments

In this section, we set our baseline using a general feature pyramid network and 3D-UNet for cost volume regularization. We then combined our proposed modules and demonstrated their effectiveness with different combinations of modules in Table 9, left. We also conducted ablation experiments for ADRP and ADIA in real scenes (TnT) in Table 9, right.

Effectiveness of ADRP: As shown in Table 9(left), we can observe that the adaptive all-pixel depth range obtained by our ADRP can effectively improve the performance on DTU dataset (Comp: from 0.3982 to 0.3342). As shown in Table 9(right), we can observe that the Recall is improved from 72.63 to 79.90 with ADPR on the Tanks and Temples intermediate benchmark. These results validate the effectiveness of our proposed ADRP.

Effectiveness of ADIA: As shown in Table 9(left), we can observe that the adaptive pixel-wise depth interval reallocated by our ADIA module can significantly improves the Acc from 0.3667 to 0.2924 on DTU dataset. The significant enhancement in performance can be attributed to the utilization of Z-score distribution-based offset calculation. In the case of small-scale scenes, the two-stage ADIAs effectively explores precise depth values within a reduced full-pixel depth range, leading to a higher level of accuracy in the depth map. Furthermore, the results exhibited in Table 9(right), demonstrate an improvement in Precision from 45.48 to 50.42 on the Tanks and Temples intermediate benchmark. By leveraging the accurate all-pixel depth range predicted by ADPR for large-scale scenes, our model demonstrates an improvement

Table 11

The quantitative experiment of two different distributions in ADIA.

Methods	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓
ADIA + Z-score distribution	**0.2924**	**0.3342**	**0.3133**
ADIA + Linear distribution	0.3345	0.3481	0.3413

in Precision, even with the utilization of two-stage ADIAs. These results corroborate the effectiveness of our proposed ADIA.

Effectiveness of ASPFNet: Table 9(left) shows that ASPFNet uses multi-path dilated convolutions to extract multi-scale contextual features, resulting in a larger receptive field and improved reconstruction quality (Acc: 0.3143 to 0.2924 and Comp: 0.3532 to 0.3342) on DTU dataset.

4.8. High-level analysis

In this section we have provided some ablation experiments for high-level analysis of the proposed ADRP module and ADIA module.

High-Level Analysis of ADRP: To verify the superiority of the image-level uniform adjustment of our proposed ADRP, we have replaced the ADRP with ADIA in stage 2 to execute an ablation study. From Table 10, we can observe that replacing ADRP with ADIA in stage 2 achieves similar Acc results (0.2901 vs. 0.2924) on the DTU benchmark, but it performs slightly worse in terms of Comp and Overall indicators compared to ARAI-MVSNet (Comp: 0.3402 vs. 0.3342, Overall:0.3152 vs. 0.3133). While for the TnT metric, ARAI-MVSNet achieves better performance in all metrics (F_1-score: 61.09 vs. 58.46, Precision: 50.42 vs. 48.11, Recall: 79.90 vs 74.48). These indicate the superiority of our ADRP, namely, predicting a reliable and accurate all-pixel depth range of the scene as a base for subsequent ADIA is effective, especially in large-scale scenes. As for the similar Acc on DTU benchmark, the reason is that targets in large scale scales have varying sizes which cannot make full use of an initial setting all-pixel depth range. While in close-range scale, the targets tend to have similar sizes, which are more likely to make full use of an initial setting all-pixel depth range.

High-Level Analysis of ADIA: To verify the superiority of the Z-score distribution-based offset calculation of our ADIA, we have conducted an ablation study by replacing the Z-score distribution in ADIA with a simpler linear distribution (i.e. removing σ from Eq. (8)). As shown in Table 11, we can observe that ADIA performs better in terms of Acc and Comp when using the Z-score distribution compared to the Linear distribution, especially in Acc (0.2924 vs. 0.3345). We attribute this improvement to the Z-score, which allows for the allocation of more depth hypothesis planes around potential ground truth depth values, which enables the estimation of more accurate depth values. As shown
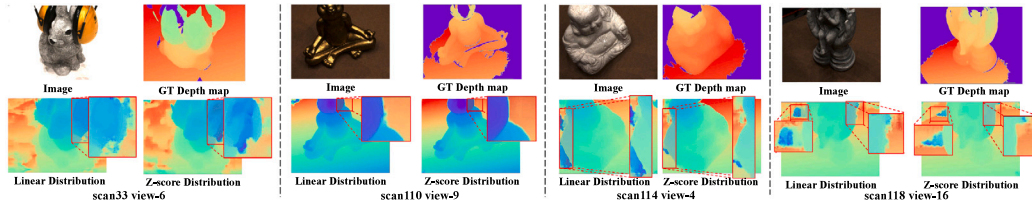
Fig. 6. Quantitative results. We compared the details of depth maps from four scans obtained by Linear distribution and Z-score distribution separately.

in Eq. (8), the value of σ is dynamically adjusted as the network continues to optimize, facilitating adaptive and high-precision adjustment of the depth plane positions for accurate depth estimation. Fig. 6 presents a qualitative comparison between ADIA+Linear distribution and ADIA+Z-score distribution for the depth map of a view from four scans on the DTU benchmark. The depth maps estimated using Z-score distribution exhibit superior overall quality compared to those obtained with Linear distribution. Particularly, examining the red boxes in the four scenes reveals that our method not only achieves better depth maps in terms of overall image quality but also provides more accurate depth estimation in detail. This improvement is attributed to the Z-score distribution, which assigns more depth hypothesis planes around potential depth values, thanks to the dynamic σ, leading to more precise depth estimation.

5. Limitations and discussion

Although our model achieves better or comparable performance than most of the state-of-the-art methods on the four benchmarks [23–26], it has several limitations: (1) Most extensive multi-view stereo (MVS) datasets, such as DTU and TnT, rely on numerous overlapped views for high-quality point cloud reconstruction. However, the performance of our model may decrease with fewer views, leading to poorer reconstruction results. This problem can be alleviated by integrating a module into MVSNet that can generate synthetic views based on limited real views, such as the neural radiance fields. (2) In current large-scale practical scene datasets, such as TnT, camera parameters are often derived from real outdoor scenes, and may not be precisely accurate. While MVSNet uses differentiable homography warping based on the camera parameters to construct the cost volume, and inaccurate camera parameters may lead to lower depth map estimation quality. To address this issue, the incorporation of a module capable of optimizing the camera parameters of inaccurate views may provide accurate camera parameters, which in turn can enable us to solve the dense structure from motion (SfM) problem. (3) With the emergence of advanced modules like Transformers, certain Transformer-based methods have gained prominence, such as MVSFormer [34]. While they exhibit superior performance in specific indicators, such as F_1-score on the TnT, this advantage comes at the cost of increased computational and memory requirements. Despite this, our method outperforms others in terms of Recall. In the future, we consider integrating these modules into more sophisticated multi-stage baselines but with efficiency into consideration. This integration is expected to yield improved or competitive performance across various metrics efficiently.

6. Conclusion

In this paper, we propose a novel multi-stage coarse-to-fine framework ARAI-MVSNet for high-quality reconstruction. Our ADRP module leverages the reference image and depth map of the previous stage to compute range adjustment parameters to achieve adaptively depth range adjustment, which can effectively zoom in an accurate all-pixel depth range of the scene. Further, the ADIA module can reallocate the pixel-wise depth interval distribution to obtain more accurate depth values by utilizing the Z-score distribution for the

adaptive variable interval partition. The experimental results demonstrate that our proposed method achieves state-of-the-art performance and exhibits competitive generalization ability compared to all listed methods. Specifically, our approach outperforms other pioneer works on the DTU dataset and obtains the highest recall score and F_1-score on the Tanks and Temples intermediate and advanced datasets. Additionally, our method achieves the lowest e_1 and e_3 scores on the BlendedMVS dataset and the highest accuracy and F_1-score on the ETH 3D dataset. These significant improvements confirm the advantages of our proposed modules. In the future, we plan to explore (1) combining ARAI-MVSNet with neural radiance fields to enable high-precision reconstruction with fewer views in MVS datasets, (2) incorporating dense bundle adjustment to jointly optimize camera poses and depth maps to enhance the robustness and accuracy of ARAI-MVSNet.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patcog.2023.109885.

References

[1] Z. Song, R. Chung, Nonstructured light-based sensing for 3D reconstruction, Pattern Recognit. (2010).

[2] B. Fang, G. Mei, X. Yuan, L. Wang, Z. Wang, J. Wang, Visual SLAM for robot navigation in healthcare facility, Pattern Recognit. (2021).

[3] Y. Fan, Q. Zhang, Y. Tang, S. Liu, H. Han, Blitz-SLAM: A semantic SLAM in dynamic environments, Pattern Recognit. (2022).

[4] S. Galliani, K. Lasinger, K. Schindler, Massively parallel multiview stereopsis by surface normal diffusion, in: ICCV, 2015, pp. 873–881.

[5] Y. Yao, Z. Luo, S. Li, T. Fang, L. Quan, Mvsnet: Depth inference for unstructured multi-view stereo, in: ECCV, 2018, pp. 767–783.

[6] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, L. Quan, Recurrent MVSNet for high-resolution multi-view stereo depth inference, in: CVPR, 2019.

[7] J. Yan, Z. Wei, H. Yi, M. Ding, R. Zhang, Y. Chen, G. Wang, Y.-W. Tai, Dense hybrid recurrent multi-view stereo net with dynamic consistency checking, in: ECCV, Springer, 2020.

[8] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, P. Tan, Cascade cost volume for high-resolution multi-view stereo and stereo matching, in: CVPR, 2020.

[9] J. Yang, W. Mao, J.M. Álvarez, M. Liu, Cost volume pyramid based depth inference for multi-view stereo, in: CVPR, 2020.

[10] S. Cheng, Z. Xu, S. Zhu, et al., Deep stereo using adaptive thin volume representation with uncertainty awareness, in: CVPR, 2020.

[11] Z. Shen, Y. Dai, Z. Rao, Cfnet: Cascade and fused cost volume for robust stereo matching, in: CVPR, 2021, pp. 13906–13915.

[12] E.I. Altman, et al., ZETATM analysis A new model to identify bankruptcy risk of corporations, J. Bank. Financ. (1977).

[13] Y. Furukawa, C. Hernández, et al., Multi-view stereo: A tutorial, Found. Trends® Comput. Graph. Vis. (2015).

[14] K.N. Kutulakos, S.M. Seitz, A theory of shape by space carving, Int. J. Comput. Vis. (2000).

[15] B. Curless, M. Levoy, A volumetric method for building complex models from range images, in: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, 1996, pp. 303–312.
[16] M. Lhuillier, L. Quan, A quasi-dense approach to surface reconstruction from uncalibrated images, IEEE TPAMI (2005).
[17] N.D. Campbell, G. Vogiatzis, Hernández, Using multiple hypotheses to improve depth-maps for multi-view stereo, in: ECCV, 2008.
[18] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, et al., Real-time visibility-based fusion of depth maps, in: ICCV, 2007.
[19] J.L. Schönberger, E. Zheng, J.-M. Frahm, M. Pollefeys, Pixelwise view selection for unstructured multi-view stereo, in: ECCV, 2016.
[20] Z. Wei, Q. Zhu, C. Min, Y. Chen, G. Wang, Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network, in: CVPR, 2021.
[21] Y. Li, Z. Zhao, J. Fan, ADR-MVSNet: a cascade network for 3D point cloud reconstruction with pixel occlusion, Pattern Recognit. (2022).
[22] S. Song, K.G. Truong, Kim, Prior depth-based multi-view stereo network for online 3D model reconstruction, Pattern Recognit. (2023).
[23] R.R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, H. Aanæs, Large scale multi-view stereopsis evaluation, in: CVPR, 2014.
[24] A. Knapitsch, J. Park, Q.-Y. Zhou, V. Koltun, Tanks and temples: Benchmarking large-scale scene reconstruction, ACM ToG (2017).
[25] Y. Yao, Z. Luo, S. Li, et al., Blendedmvs: A large-scale dataset for generalized multi-view stereo networks, in: CVPR, 2020.
[26] T. Schops, J.L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, A. Geiger, A multi-view stereo benchmark with high-resolution images and multi-camera videos, in: CVPR, 2017.
[27] F. Wang, S. Galliani, C. Vogel, P. Speciale, M. Pollefeys, Patchmatchnet: Learned multi-view patchmatch stereo, in: CVPR, 2021.
[28] X. Ma, Y. Gong, Q. Wang, et al., EPP-MVSNet: Epipolar-assembling based depth prediction for multi-view stereo, in: CVPR, 2021.
[29] F. Wang, S. Galliani, C. Vogel, M. Pollefeys, IterMVS: Iterative probability estimation for efficient multi-view stereo, in: CVPR, 2022.
[30] R. Peng, R. Wang, Z. Wang, et al., Rethinking depth estimation for multi-view stereo: A unified representation, in: CVPR, 2022.
[31] J. Yang, J.M. Alvarez, M. Liu, Non-parametric depth distribution modelling based depth inference for multi-view stereo, in: CVPR, 2022.
[32] H. Liu, X. Tang, S. Shen, Depth-map completion for large indoor scene reconstruction, Pattern Recognit. (2020).
[33] Y. Ding, W. Yuan, Q. Zhu, et al., Transmvsnet: Global context-aware multi-view stereo network with transformers, in: CVPR, 2022.
[34] C. Cao, X. Ren, Y. Fu, MVSFormer: Multi-view stereo by learning robust image features and temperature-based depth, TMLR (2023).
[35] J. Liao, Y. Ding, Y. Shavit, D. Huang, S. Ren, et al., Wt-mvsnet: window-based transformers for multi-view stereo, in: NIPS, 2022.

Song Zhang received the BEng degree with honors from Anhui University, China, in 2020. Currently, he is pursuing a Ph.D. degree at the University of the Chinese Academy of Sciences. His research interests focused on 3D point cloud reconstruction, Vision Transformer, and image processing.

Wenjia Xu received the BEng degree with honors from the Beijing Institute of Technology in 2016, and the Ph.D. degree from the University of Chinese Academy of Sciences in 2022. She worked as a visiting Ph.D. student at the MaxPlanck Institute for Informatics in Germany between 2019–2020. Currently, she is a research associate professor at the Beijing University of Posts and Telecommunications. Her research interest includes learning with limited supervision for computer vision tasks and remote sensing images, and explainable machine learning.

Zhiwei Wei received a B.S. degree and Ph.D. degrees in GIS from Wuhan University, China, in 2015 and 2020. He is currently an Assistant Professor at the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include creating explanatory and exploratory animations for spatial data, automated understanding, and processing of spatial data via spatial knowledge graph, and 3D Reconstruction for Enhanced VR.

Lili Zhang received a B.S. degree from Northwestern Polytechnical University, China, and a Ph.D. degree from the University of Chinese Academy of Sciences in 2022. She is currently an Associated Professor and a master advisor Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include semantic segmentation and image processing.

Yang Wang received a Ph.D. degree in cartography from Pe-king University, China, in 2014. She is currently an Associated Professor and a master advisor Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her research interests include the analysis and visualization of temporal and geographical information.

Junyi Liu received a B.S. degree and Ph.D. degrees from the National University of Defense Technology, China, in 1995 and 2001. She is currently a Professor and a doctoral supervisor Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests focused on network information systems, multi-source information fusion, and image information processing.