A Global Depth-Range-Free Multi-View Stereo Transformer Network with Pose Embedding

Yitong Dong[1*]   **Yijin Li**[1*]   **Zhaoyang Huang**[2]   **Weikang Bian**[2]
Jingbo Liu[1]   **Hujun Bao**[1]   **Zhaopeng Cui**[1]   **Hongsheng Li**[2]   **Guofeng Zhang**[1†]
[1]State Key Lab of CAD&CG, Zhejiang University   [2]CUHK MMLab

Appendix

In this document, references that point to the main manuscript will be referenced as "P-".

A Implementation Details

A.1 Epipolar Line Searching Range

Given a pixel p_r in reference image I_0, we can determine the position p_s in the source image I_i by Eq. 1 based on multi-view geometric consistency.

$$K_i^{-1} p_s d_s = R \cdot \left(K_0^{-1} p_r d_r \right) + T, \tag{1}$$

where d_r denotes the depth in reference view, d_s denotes the depth in source view. R and t denote the rotation and translation between the reference and the source view, K_0^{-1} and K_i^{-1} denote the intrinsic matrices. In order to determine the range of epipolar line matching the observability condition $d_r > 0, d_s > 0$ in multi-view images, we segregate d_r, d_s and treat them individually. Specifically, to determine the relationship between the depth in the reference view d_r and the coordinates in the source view p_s, we multiply both sides of the equation by a skew-symmetric matrix of $K_i^{-1} p_s$:

$$(K_i^{-1} p_s)^{\wedge} \cdot K_i^{-1} p_s d_s = (K_i^{-1} p_s)^{\wedge} \cdot R \cdot \left(K_0^{-1} p_r d_r \right) + (K_i^{-1} p_s)^{\wedge} \cdot T, \tag{2}$$

where $(\cdot)^{\wedge}$ represents the skew-symmetric matrix function:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{\wedge} = \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix}, \tag{3}$$

Due to the inherent properties of the skew-symmetric matrix $x^{\wedge} \cdot x = 0$, we can transform Eq. 2 into an equation between d_r and p_s.

$$(K_i^{-1} p_s)^{\wedge} \cdot R \cdot \left(K_0^{-1} p_r d_r \right) + (K_i^{-1} p_s)^{\wedge} \cdot T = 0, \tag{4}$$

Let $(K_i^{-1} p_s) = (p_{sx}, p_{sy}, p_{sz})^T$, $R \cdot K_0^{-1} p_r = (p_{rx}, p_{ry}, p_{rz})^T$ and $T = (t_x, t_y, t_z)^T$, we can obtain the reference depth and source depth by Eq. 5.

$$\begin{cases} p_{sx} = (p_{rx} d_r - t_x)/(p_{rz} d_r - t_z) \\ p_{sy} = (p_{ry} d_r - t_y)/(p_{rz} d_r - t_z) \end{cases} \tag{5}$$

---

[*]Equal contribution
[†]Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).
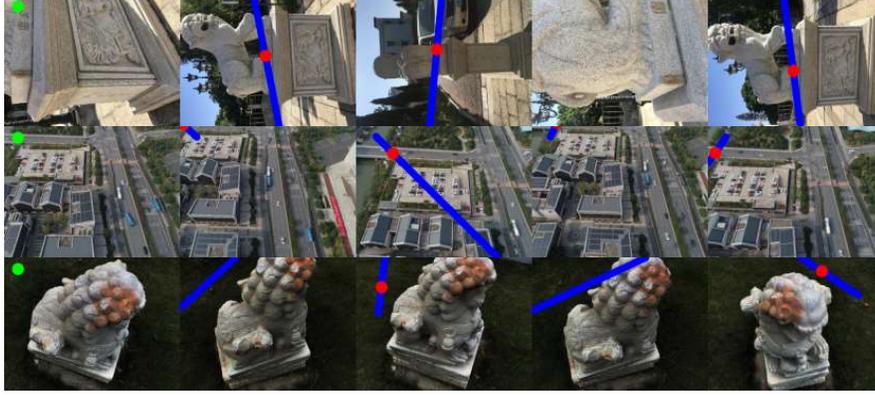
Figure 1: Epipolar line searching range on DTU [1] and BlendedMVS [2] datasets. Green points represent the pixel on the reference image, red points represent the pixel on the source images and blue lines represent the epipolar line. Due to physical constraints of positive depth for both the reference image and source image, the search range along the epipolar line of the source image is narrowed down to a specific interval.

We can then obtain the two endpoints on the epipolar line in the source image, thereby defining the search range along the epipolar line:

$$
\begin{cases}
\lim_{d_r \to 0} p_s = \lim_{d_r \to 0} K_i \cdot \begin{pmatrix} p_{sx} \\ p_{sy} \\ 1 \end{pmatrix} = K_i \cdot \begin{pmatrix} t_x/t_z \\ t_y/t_z \\ 1 \end{pmatrix} \\
\lim_{d_r \to +\infty} p_s = \lim_{d_r \to +\infty} K_i \cdot \begin{pmatrix} p_{sx} \\ p_{sy} \\ 1 \end{pmatrix} = K_i \cdot \begin{pmatrix} p_{rx}/p_{rz} \\ p_{ry}/p_{rz} \\ 1 \end{pmatrix}
\end{cases}
\tag{6}
$$

We employ a similar approach to determine the epipolar searching range that satisfies the observability condition $d_s > 0$ in the source view. It is crucial to ensure that both the source depth d_s and the reference depth d_r are greater than zero. Therefore, the final sampling interval is determined by the intersection of the two intervals. The initial position p_s^0 of the source image is determined by selecting the midpoint of the search range along the epipolar line. Fig. 1 illustrates the epipolar searching range for a pixel in the reference image, corresponding to multi-view source images from the DTU [1] and BlendedMVS [2] datasets. It can be observed that leveraging the epipolar line ranges derived from physical properties effectively narrows the optimization region, resulting in more accurate sampling.

A.2 Depth Normalization

DispMVS [3] utilizes the depth range in the depth regression process to enhance the stability of the network. Specifically, the depth range is employed for initialization in Eq. 7, and regularization for depth upsampling and fusion in Eq. 8.

$$
\frac{1}{d_r^0} = rand \cdot \left(\frac{1}{d_{min}} - \frac{1}{d_{max}} \right) + \frac{1}{d_{max}},
\tag{7}
$$

$$
d_{norm} = \frac{1/d - 1/d_{max}}{1/d_{min} - 1/d_{max}},
\tag{8}
$$

In the initialization, DispMVS ensures the reliability of d_0 by randomly selecting an initial position within the depth range, thereby facilitating easier convergence of the network. In the multi-view depth fusion, DispMVS converts the epipolar disparity flow to depth by triangulation and utilizes depth normalization to decrease the effect of outliers.

To further enhance the network's generalization ability regarding depth range, we attempt to eliminate depth range-related operations within the network. Specifically, this entails removing certain initialization and depth normalization procedures. Thanks to the previously computed epipolar line range, we can obtain a relatively accurate initial value. Additionally, with the introduction of the Multi-view

2

Table 1: Quantitative evaluation towards VideoFlow setting on DTU validation set. The lower the Endpoint Error (EPE), e2, e4, e8, the better.

Method	EPE(mm) ↓	e2(mm) ↓	e4(mm) ↓	e8(mm)
Concatenation Setting [3]	3.376	0.139	0.070	0.041
Ours	**3.298**	**0.123**	**0.063**	**0.038**

Disparity Attention (MDA) module in our network, we can better utilize geometric information across multi-view frames, leading to improved convergence. Consequently, we set the initial position p_s^0 in the reference image as the midpoint of the epipolar line searching range (p_s^{min}, p_s^{max}) in Eq. 9 and then obtain the initial position p_s^0.

$$p_s^0 = (p_s^{min} + p_s^{max})/2, \tag{9}$$

We remove the depth normalization to further reduce the reliance on the depth range and enhance the network's generalization capabilities. Experimental results demonstrate that our setup does not result in performance degradation and we can still achieve comparable reconstruction accuracy.

A.3 Feature Extraction

Given a reference image I_0 and multi-view source images $\{I_i\}_{i=1}^{N-1}$, we utilize two share-weighted feature extraction modules to extract image features $F_0^l \in \mathbb{R}^{H \times W \times C}$ and $\{F_i^l\}_{i=1}^{N-1} \in \mathbb{R}^{H \times W \times C}$. These features are extracted at two down-sampled scale levels, corresponding to the resolution of $1/4^l$, with channel dimensions C of $96,128$ and levels l of $1,2$. Additionally, we design a context feature network to extract features from the reference image I_0, preparing for subsequent GRU updates. Notably, the context network does not share weights with the feature extraction network.

B More Ablation Study

B.1 VideoFlow Setting

Our method is inspired by VideoFlow [4], which estimates optical flows for multiple frames by effectively leveraging temporal cues. However, we decided not to adopt some of the structures from VideoFlow. In this section, we provide an analysis of our decision.

VideoFlow utilizes the Global Motion Aggregation (GMA) module [5] to perceive context information from the reference image and construct global features $F_g \in \mathbb{R}^{H \times W \times H \times W}$. The GMA module utilizes a transformer to identify long-range dependencies among pixels in the reference image and conducts global aggregation on the associated motion features. Therefore, the input to the GRU update module in VideoFlow consists of the motion feature F_m, along with the global context feature F_g and the hidden state feature h_k, where k representing the $k-th$ iteration. However, due to the typically high resolution with 864×1152 of the test dataset in the MVS task, which results in significant memory consumption during GMA computation, we have decided to abandon the GMA module.

Additionally, VideoFlow concatenates the flow and correlation of all images along the channel dimension, and then the concatenated features are input into the motion encoder. The flow $F_k \in \mathbb{R}^{B \times N \times H \times W}$ of all images are directly output through the GRU module. In our approach, we input the respective optical flow and correlation of each source image into the disparity encoder separately. Subsequently, after obtaining the disparity features, we fuse the multi-view information using the Multi-view Disparity Attention (MDA) module. Table 1 demonstrates that our method can more efficiently integrate the potential relationships between 2D epipolar disparity flows across multi-view frames. EPE (Endpoint Error) represents the average $l - 1$ distance between ground truth and the prediction within the mask region, $e2$, $e4$ and $e8$ represent the proportions of pixels with depth error larger than 2, 4, and 8, respectively.

B.2 Multi-view Loss

In this section, we explore the integration of multi-view constraints into the loss function as regularization terms. The objective is to prompt the network to learn consistent 3D information across

3

Table 2: Quantitative evaluation towards multi-view loss on DTU validation set. The lower the Endpoint Error (EPE), e2, e4, e8, the better.

Loss	EPE(mm) ↓	e2(mm) ↓	e4(mm) ↓	e8(mm) ↓
L1 loss add multi-view loss [3]	3.322	0.125	0.064	0.038
L1 loss	**3.298**	**0.123**	**0.063**	**0.038**

Table 3: Ablation study on number of input views N on DTU evaluation set [1]. The lower the Accuracy (Acc), Completeness (Comp), Overall, the better.

DTU	3 views	4 views	5 views	6 views	7 views	8 views	9 views	10 views
EPE(mm)	5.375	5.108	4.992	4.964	4.929	4.967	5.002	5.024

multi-view source images. To accomplish this, we compute the variance of depths obtained from multiple frames and incorporate it into the loss function.

Table 2 suggests that imposing such constraints aggressively introduces unknown biases into the network, resulting in diminished performance compared to direct utilization of L1 loss. Particularly in scenarios where some images experience severe occlusion or objects extend beyond the field of view, enforcing consistency among depths corresponding to multiple frames becomes impractical and may even introduce noise.

B.3 Dense View Setting

For images with a resolution of 512x640, our method can handle up to 77 views in V100 during testing phase, and up to 5 views simultaneously during training phase. Using the model trained on the DTU dataset with 5 views, we verify the impact of the number of views on system accuracy. Due to DTU providing co-visible relationships for only 10 frames, we selected 10 frames as the upper limit. The results are shown in Table 3, which uses 2D metric. It can be observed that the error decreases as the number of used views increases and gradually basically coverage. Increasing the view during training may further increase the performance after convergence.

B.4 Initializations of Depth

We add the comparison with different initializations of depth. However, due to the unknown depth range, it is not feasible to design a random sampling range for 3D sampling, which is impractical. Additionally, when the initial depth is set to 0, significant noise can occur during feature warping. Therefore, we designed three sets of ablation experiments in Table 4: setting all depths to one, randomly initializing within the epipolar search range, and initializing fixed at the left endpoint of the epipolar line.

C   More Experiment Results

C.1   Evaluation on BlendedMVS Dataset

We evaluate the proposed method on the evaluation set of BlendedMVS dataset [2]. Following [6], we test any method with five input images on the standard test set, composed of 7 heterogeneous scenes. We compute the mean absolute error (MAE), root mean squared error (RMSE), and the percentage of pixels having depth error larger than a given threshold ($> \tau$). As shown in Table 5, our method has the best overall score among depth-range-free methods.

Table 4: Comparison of Initializations

Ablation	Acc.(mm)↓	Comp.(mm)↓	Overall(mm)↓
The Left Endpoint	0.389	4.373	2.381
Randomly 2D Initializing	2.597	2.211	2.404
Ours	**0.338**	**0.331**	**0.335**

4

Table 5: Quantitative point cloud evaluation results on BlendedMVS evaluation set.

Method	MAE ↓	RMSE ↓	> 1 ↓	> 2 ↓	> 3 ↓	> 4 ↓	> 8 ↓
MVSNET [7]	0.6168	1.5943	0.1392	0.0731	0.0457	0.0309	0.0103
R-MVSNet [8]	0.7815	1.7397	0.1864	0.1007	0.0637	0.0433	0.0141
PatchmatchNet [9]	0.3849	1.3581	0.0749	0.0386	0.0247	0.0175	0.0067
Vis-MVSNet [10]	**0.3318**	**1.2396**	**0.0662**	**0.0323**	**0.0197**	**0.0133**	**0.0044**
IterMVS [11]	**1.9644**	**9.0190**	0.2322	0.1384	0.0989	0.0767	0.0392
CER-MVS [12]	2.1666	26.934	**0.0752**	**0.0441**	**0.0316**	**0.0247**	**0.0138**
RAMDepth [6]	0.2982	1.1724	0.0645	0.0285	0.0159	0.0102	0.0033
DispMVS [3]	0.2734	0.9471	0.0593	0.0277	0.0161	0.0104	0.0033
Ours	**0.2531**	**0.8999**	**0.0545**	**0.0250**	**0.0143**	**0.0092**	**0.0030**
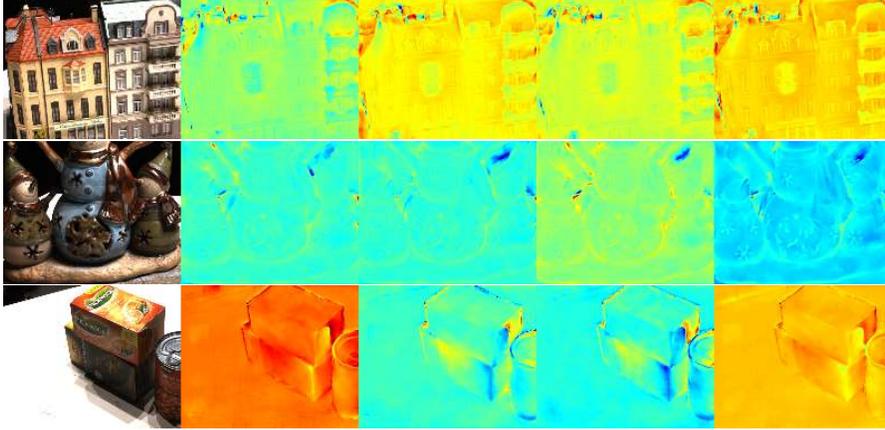
Figure 2: The epipolar disparity flows across multi-view

C.2 Depth Range

As shown in Fig. P-1, since we have completely removed the depth range prior, the output results are not affected by any errors in the depth range. In contrast, IterMVS samples based on the depth range, which prevents it from obtaining the correct depth for foreground (425mm - 552.5mm) and background (807.5mm - 935mm) points. DispMVS's initialization also rely on the depth range. During each iteration, it uses the depth range to apply a depth normalization and filter out outliers for stability, so an underestimated depth range can significantly affect its performance.

C.3 Visualization of Epipolar Disparity Flows

To consider the information interaction of multiple source images during 2D sampling, it is necessary to train the flow. As shown in Fig. 2, after the MDA and GRU modules, the flow obtained from different source images is consistent in terms of details. By incorporating geometric information, although the flow magnitudes on different source images vary, the representation of edges and other details is unified.

C.4 Visualization of GRU Updating

As shown in Fig. 3 (b), the depth error decreases progressively with each iteration. The vertical axis represents the depth error, and the horizontal axis represents the number of iterations. Iterations 0-7 correspond to the coarse stage, while iterations 8-9 correspond to the fine stage. Fig. 3 (a) shows the depth maps on DTU, in which we can see that the depth map recovers from coarse to fine.

C.5 2D Experiment Results

As illustrated in Fig. 4, our method achieves more accurate 2D depth maps on the DTU and BlendedMVS datasets compared to DispMVS [3].

5
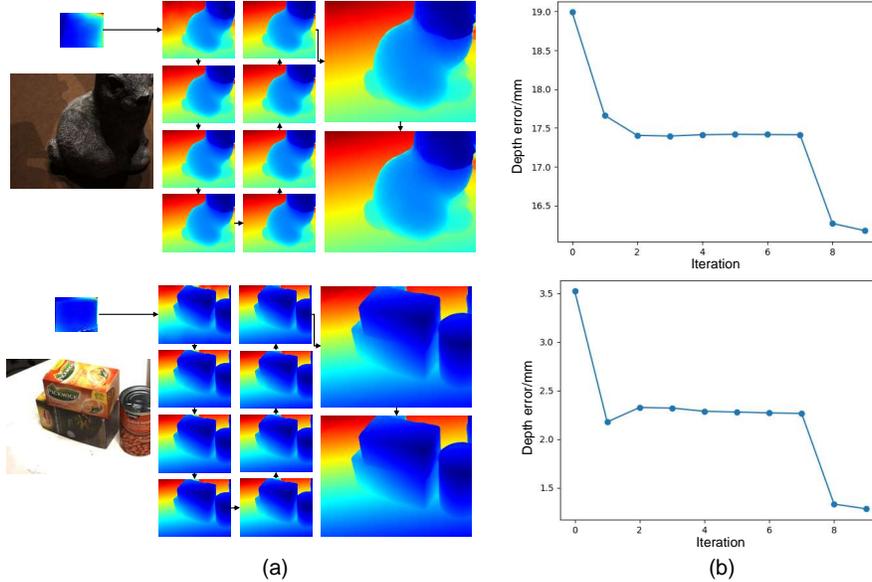
Figure 3: Depth Maps in each iteration(a) and Depth Error variation with GRU iterations(b).

C.6 Generalization Capability

To evaluate the generalization of our method, we used an iPad to capture data and add inference experiments in multiple scenes. We collected images from various real-world environments, with intrinsic and extrinsic parameters obtained by running COLMAP. As shown in Fig. 5, our model is capable of generating dense point cloud reconstructions for both indoor and outdoor environments. However, the accuracy of these reconstructions is somewhat lacking, which is affected by the following three factors:

Inaccurate Camera Pose: The pose estimation from COLMAP is far from satisfying. In contrast, the evaluation benchmark provides more accurate camera pose. For example, DTU uses a structured light scanner and MATLAB calibration toolbox for camera pose estimation. The inaccurate camera pose can lead to large error during the MVS process.

Image Quality Issues: As illustrated in Fig. 6, issues such as overexposed, inadequate lighting and blurriness affect the reconstruction quality. These deficiencies in image quality contribute to the observed inaccuracies in the point cloud.

Training Data Limitations: Our model was trained on the DTU dataset, which is relatively small and features a narrow range of scenes. While our model can effectively mitigate depth range effects in various environments, it still struggles with fine detail accuracy. The limited diversity of the DTU dataset constrains the model's ability to capture detailed features accurately. Constructing an MVS dataset with diverse scenes is a promising approach to enhance the robustness and accuracy of point cloud reconstructions.

C.7 Larger-scale Experiments

We collected some real-world data to test our method in more large-scale environment and also test its generalization. The camera intrinsic and camera pose are both obtained by running COLMAP. As shown in Fig. 5, our model is capable of generating dense point cloud reconstructions for the collected data, which shows basic generalization ability. However, the accuracy of these reconstructions is somewhat lacking. We guess one of the primary factors is the inaccurate camera pose from the COLMAP. It is a promising direction to explore the joint optimization of the camera pose for the MVS in the future. Besides, the limited training data also hinders the performance of our method. Compared with Dust3r which is trained with a mixture of eight datasets, covering millions of image data, our method and other MVS methods are only trained on DTU or BlendedMVS. How to utilize these large datasets for training MVS is also one of our future work.

6
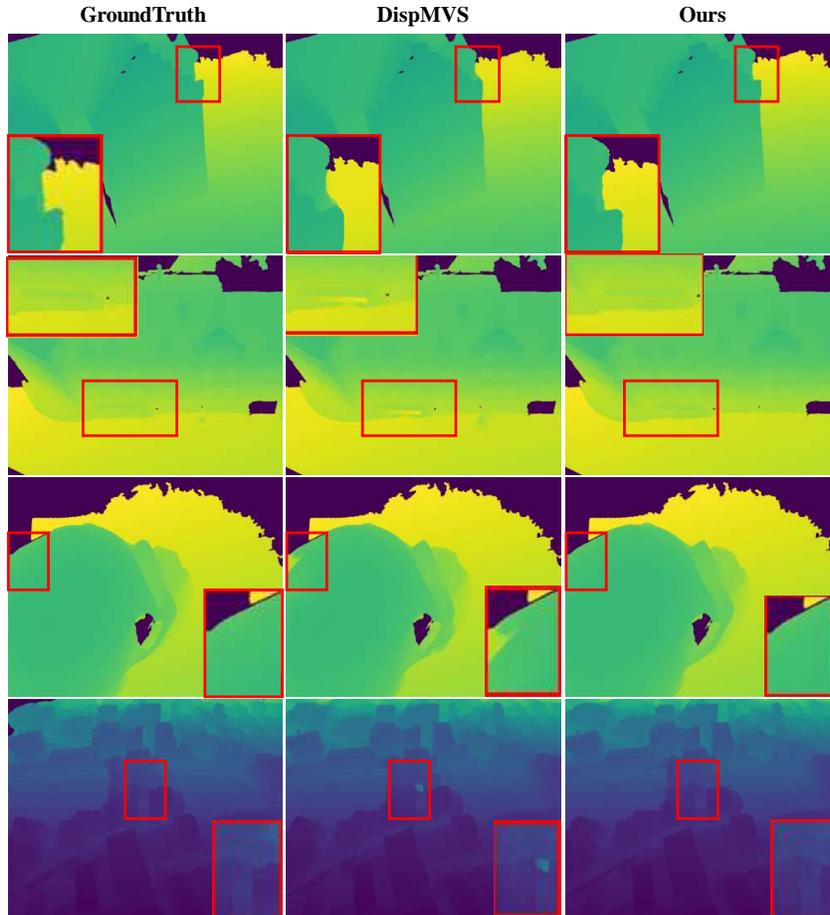
Figure 4: The qualitative comparison on DTU [1] and BlendedMVS [2] datasets. Our method achieves more accurate depth results, especially along the object boundary.

Figure 5: Visualization of Point Clouds of Custom Data.

C.8   Known Depth Prior

To validate the effectiveness of the depth prior, we designed the experiment. The initial depth is reverted to random initialization based on the depth prior, similar to DispMVS. The results are shown in Table 6. It can be observed that adding depth prior can improve performance to some extent, but the difference compared to our depth-range-free method is not significant. This indicates that the current initial point selection strategy and the design of the Transformer enable the network to regress to the correct range, resulting in accurate depth estimation.

7

Figure 6: Challenges of Custom Data in MVS Point Cloud Reconstruction.

Table 6: Quantitative evaluation of the model with depth range prior on the DTU dataset. The lower the Accuracy (Acc), Completeness (Comp), Overall, the better.

Ablation	Acc.(mm)↓	Comp.(mm)↓	Overall(mm)↓
Random Depth Initialization among Depth Range Prior	**0.331**	0.335	**0.333**
Ours	0.338	**0.331**	0.335

D Limit

Compared to other methods that uniformly sample depth based on a depth range, our network requires a more powerful retrieval capability to regress the correct depth due to the lack of depth range priors and outperform known depth-free methods. There are two main reasons for the discrepancy between our method and cost-volume approaches.

One reason is the search gratuity. Although our method addresses the receptive field problem to a large extent by sampling points along the epipolar line at features with different scales, iterating over the depth range of significantly increases the search gratuity compared to uniformly sampling within a predefined depth range.

The second reason is the development potential of datasets for deep learning network. Existing MVS datasets, such as DTU, have a relatively uniform depth distribution, mostly around a mean depth of 600. This allows methods directly based on a narrow predefined depth range to achieve precise convergence, limiting the advantage of our method. However, in real-world scenarios, there are many scenes with a wide depth distribution, such as near and far objects, where the background cannot be crudely masked out like in the DTU dataset. In such cases, depth cannot be recovered with a narrow depth prior, necessitating an expanded depth search range, which increases the difficulty of convergence inevitably. I think enhancing accuracy over a large depth range is a crucial problem that MVS must address in the future. Currently, our ongoing work focuses on optimizing the acquisition of initial values and constructing more diverse datasets to endow the network with stronger learning capabilities.

E Point Cloud Results

E.1 More Point Cloud Results

We compare our method with several state-of-the-art methods [13; 3] on DTU dataset [1], with Fig. 7 illustrating superior performance of our approach. We show more qualitative results on DTU dataset in Fig. 8 and Tanks and Temples dataset in Fig. 9.

Additionally, we use the model trained on the large-scale BlendedMVS dataset with image resolution as 576×768 and number of input images as 5 to evaluate the BlendedMVS. The BlendedMVS dataset [2] is a large-scale outdoor multi-view stereo dataset that contains a diverse array of objects

8
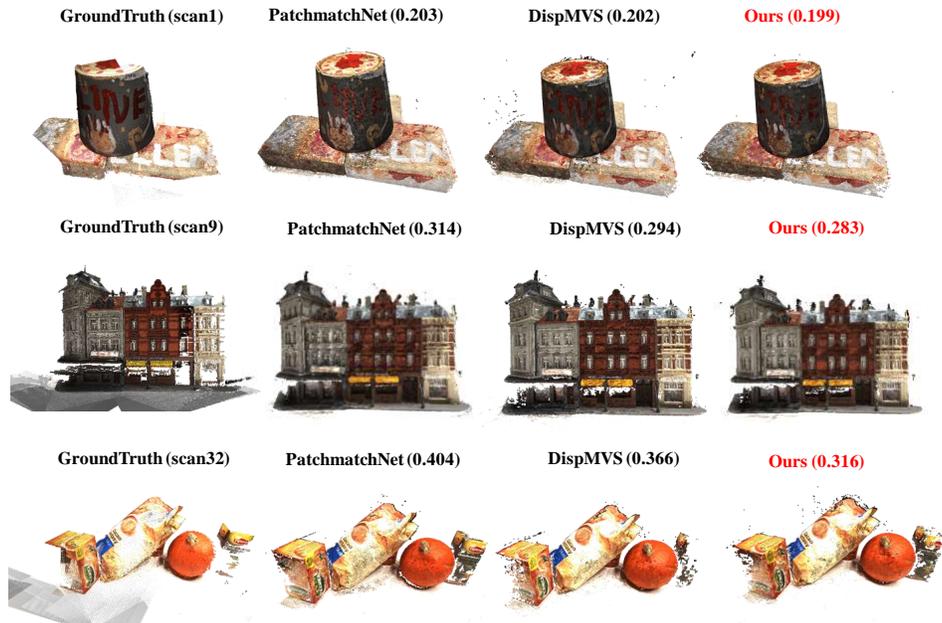
Figure 7: Comparison of reconstructed results with several state-of-the-art methods [13; 3] on DTU dataset [1]. Each method corresponds to the number in parentheses, representing the 3D overall metric (lower is better). We highlight the best-performing method in red.

and scenes. We set the image resolution as 576×768 and the number of input images as 7 for the evaluation phase of it. We showcase the performance of the point cloud on the BlendedMVS dataset in Fig. 10.

E.2 Floater Artifacts

As shown in Fig. 11, we compare our method with MVSFORMER++(SOTA) and find that point clouds inevitably exhibit artifacts in MVS task. Artifacts are generated during the depth fusion step. For point cloud fusion, we directly sampled the PCD method from DispMVS. This method selects multiple relevant depth views for each image to perform back-projection. After threshold filtering, the back-projected depth is weighted and combined with the current depth, which can lead to the generation of floater artifacts. This is due to 2D depth errors and inconsistencies across multi-view frames. Adjusting the threshold can mitigate this issue but may affect the overall quality of the 3D point cloud.

9

Figure 8: Point clouds in DTU evaluation set [1].

10
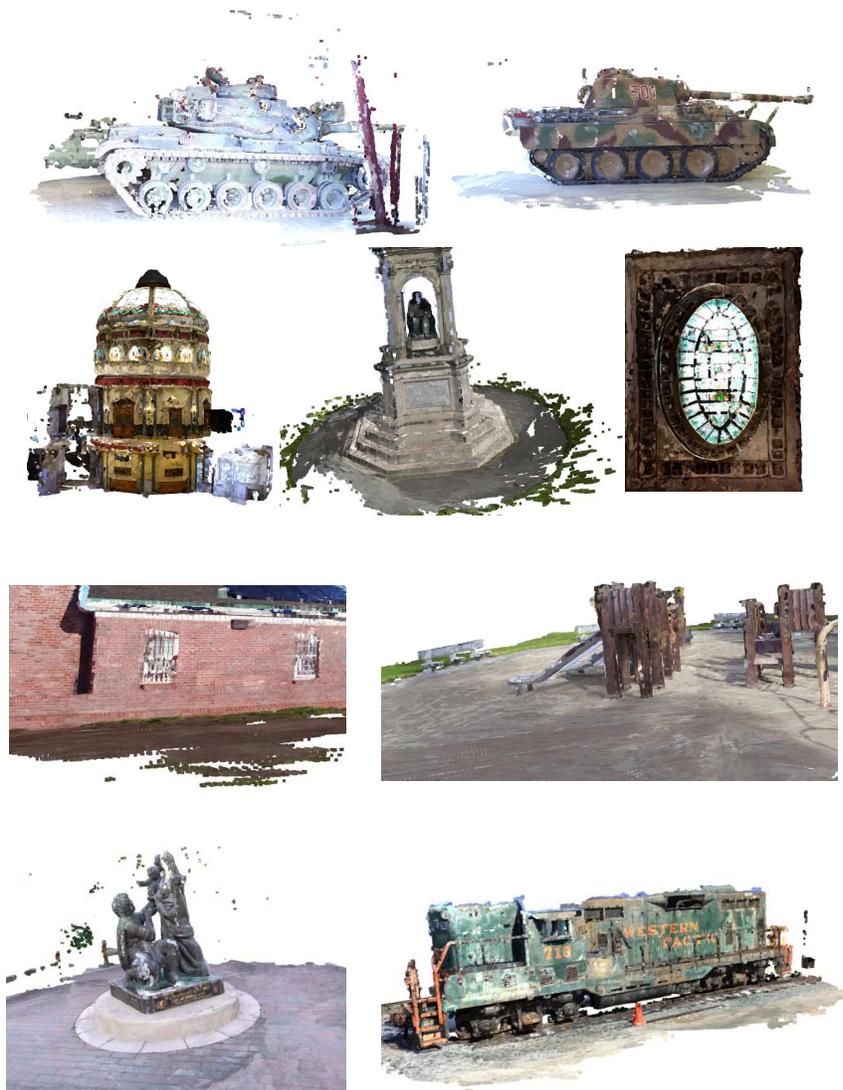
Figure 9: Point clouds of Tanks and Temples Benchmark [14].

11
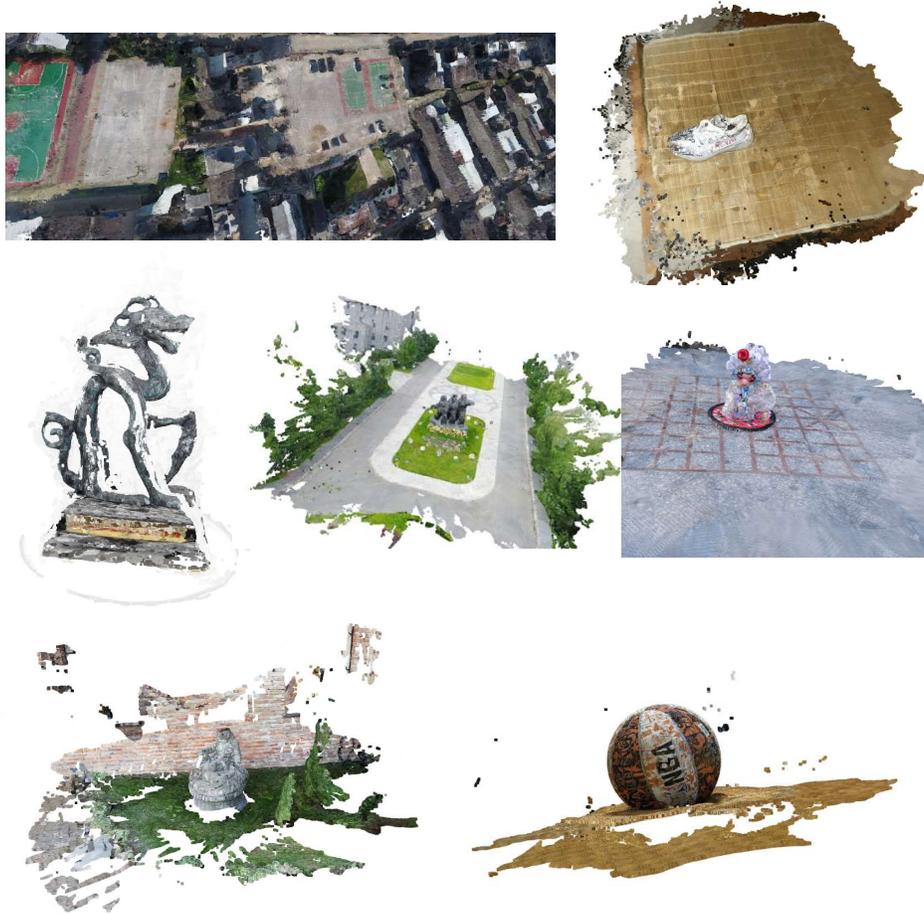
Figure 10: Point clouds in BlendedMVS validation set [2].

12

Ours   **MVSFORMER++ (SOTA)**

Figure 11: Comparison of Point Cloud Floater Artifacts.

13

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016.

[2] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] Qingsong Yan, Qiang Wang, Kaiyong Zhao, Bo Li, Xiaowen Chu, and Fei Deng. Rethinking disparity: a depth range free multi-view stereo based on disparity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3091–3099, 2023.

[4] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. *arXiv preprint arXiv:2303.08340*, 2023.

[5] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9772–9781, 2021.

[6] Andrea Conti, Matteo Poggi, Valerio Cambareri, and Stefano Mattoccia. Range-agnostic multi-view depth estimation with keyframe selection. *arXiv preprint arXiv:2401.14401*, 2024.

[7] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[8] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14194–14203, June 2021.

[10] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network, 2020.

[11] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Itermvs: Iterative probability estimation for efficient multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8606–8615, June 2022.

[12] Zeyu Ma, Zachary Teed, and Jia Deng. Multiview stereo with cascaded epipolar raft. In *European Conference on Computer Vision*, pages 734–750. Springer, 2022.

[13] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6187–6196, October 2021.

[14] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

14