RRT-MVS: Recurrent Regularization Transformer for Multi-View Stereo

Jianfei Jiang, Liyong Wang, Haochen Yu, Tianyu Hu, Jiansheng Chen, Huimin Ma*

School of Computer and Communication Engineering, University of Science and Technology Beijing, China
{jiangjf, d202410454, haochen.yu}@xs.ustb.edu.cn, {tianyu, jschen, mhmpub}@ustb.edu.cn

Abstract

Learning-based multi-view stereo methods aim to predict depth maps for reconstructing dense point clouds. These methods rely on regularization to reduce redundancy in the cost volume. However, existing methods have limitations: CNN-based regularization is restricted to local receptive fields, while Transformer-based regularization struggles with handling depth discontinuities. These limitations often result in inaccurate depth maps with significant noise, particularly noticeable in the boundary and background regions. In this paper, we propose a Recurrent Regularization Transformer for Multi-View Stereo (RRT-MVS), which addresses these limitations by regularizing the cost volume separately for depth and spatial dimensions. Specifically, we introduce Recurrent Self-Attention (R-SA) to aggregate global matching costs within and across the cost maps and filter out noisy feature correlations. Additionally, we present Depth Residual Attention (DRA) to aggregate depth correlations within the cost volume and a Positional Adapter (PA) to enhance 3D positional awareness in each 2D cost map, further augmenting the effectiveness of R-SA. Experimental results demonstrate that RRT-MVS achieves state-of-the-art performance on the DTU and Tanks-and-Temples datasets. Notably, RRT-MVS ranks first on both the Tanks-and-Temples intermediate and advanced benchmarks among all published methods.

Introduction

Multi-View Stereo (MVS) is a challenging task in computer vision, aims to recover dense 3D point clouds of scenes or objects from a series of calibrated images. Although traditional MVS methods (Galliani, Lasinger, and Schindler 2015; Xu and Tao 2020b; Xu et al. 2022a; Wang et al. 2023; Yuan et al. 2024) have shown advancements in recent years, they are hindered by manually crafted feature representations, and still struggle with issues such as reflective surfaces, occlusions, and areas with low texture. To address these challenges, researchers have harnessed the powerful capabilities of deep neural networks to develop learning-based MVS methods. These methods typically involve four key steps: feature extraction, cost volume construction, cost volume regularization, and depth prediction.

*Corresponding author

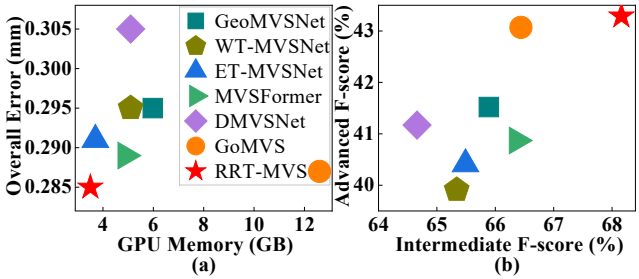

Figure 1: Comparison with state-of-the-art methods in terms of (a) overall error and GPU memory consumption on the DTU test set *(lower is better)*, and (b) generalization ability on the Tanks-and-Temples benchmark *(higher is better)*.

Learning-based MVS methods primarily focus on calculating pair-wise feature correlations between reference image and each source image by differentiable wrapping (Yao et al. 2018), and aggregate them into a cost volume. Although current MVS methods have attempted to improve the accuracy of cost volume by enhancing feature extraction capabilities (Cao, Ren, and Fu 2022; Liu et al. 2023) or cost volume construction methods (Wei et al. 2021; Wang et al. 2022c), constructing the cost volume inevitably introduces redundant background information and incorrect matching costs. Consequently, effectively tackling cost volume regularization is crucial for MVS tasks.

Recently proposed state-of-the-art MVS methods often use a cascade structure to predict depth maps in a coarse-to-fine manner. As a crucial step in cascade-based MVS methods, cost volume regularization is typically achieved using either CNNs or Transformers. CNN-based methods are constrained by their receptive fields, lacking the ability to capture the global context of the cost volume, and thus can only perform local cost volume regularization. Transformer-based methods require converting the format of 3D cost volume to be processed by Transformers. Specifically, WT-MVSNet (Liao et al. 2022) and CostFormer (Chen et al. 2023) partition the depth and spatial dimensions into smaller 3D windows and then apply window-based attention (Liu et al. 2021). MVSFormer++ (Cao, Ren, and Fu 2024) combines the depth and spatial dimensions into 1D sequences and uses vanilla attention (Dao 2023) for cost volume reg-

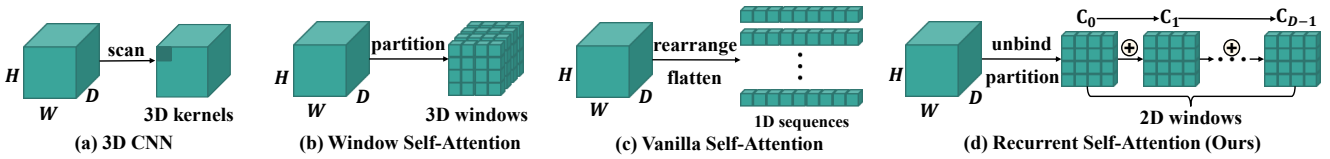

Figure 2: An illustration of different cost volume regularization methods in cascade-based MVS networks. Our proposed (d) *Recurrent Self-Attention* segments the cost volume into cost maps $\{C_i\}_{i=0}^{D-1}$ along the depth dimension D, applies 2D window-based attention for regularization, and creates connections between the cost maps across the depth dimension.

ularization. Generally, the spatial dimensions represent the 2D coordinates of pixels and are continuous across the entire scene. In contrast, the depth dimension is obtained through depth sampling, which is sparser compared to the spatial dimension and cannot cover the entire scene at all stages. Notably, neighboring 2D coordinates in images may not correspond to depth values from the same object. Simply merging the depth and spatial dimensions or dividing them into 3D windows can significantly impact depth estimation in areas with depth discontinuities, such as object boundaries or backgrounds. This problem is exacerbated in cascade structures due to the increasingly narrow depth ranges and the accumulation of depth errors, resulting in suboptimal reconstruction results.

In this paper, we propose RRT-MVS, which employs a Recurrent Regularization Transformer (RRT) for global cost volume regularization. Considering the discontinuity of the depth dimension in cascade-based MVS methods, we split the 3D cost volume into 2D cost maps along the depth dimension, with each cost map corresponding to a depth hypothesis plane. RRT combines convolutional attention and self-attention mechanisms, effectively aggregating matching cost information and filtering redundant feature correlations in the cost volume. This is achieved through the collaboration of three components, which separately regularize the cost volume in the depth and spatial dimensions. Specifically, we propose a Recurrent Self-Attention (R-SA), which uses self-attention to eliminate redundant feature correlations within each depth hypothesis plane, and convolutional attention to aggregate matching information within each cost map, maintaining continuity in the depth dimension through a recurrent mechanism. Due to the design of our recurrent architecture, R-SA avoids cubic computational complexity. To enhance the spatial 3D positional awareness of R-SA, we employ a Positional Adapter (PA) to encode 3D positional information onto 2D cost maps. Additionally, we introduce Depth Residual Attention (DRA) to aggregate information along the depth direction, effectively addressing the limitations of the recurrent mechanism. Unlike existing cascade-based MVS methods, our approach maintains independence in the depth dimension during cost volume regularization. We conducted extensive experiments to demonstrate the effectiveness and efficiency of RRT-MVS. As shown in Figure 1(a), compared with the current state-of-the-art method GoMVS (Wu et al. 2024), RRT-MVS achieves the best performance on the DTU (Aanæs et al. 2016) dataset while reducing GPU memory consumption by 72.17%. Additionally, RRT-MVS also demonstrates strong generalization capability on the Tanks-and-Temples (Knapitsch et al. 2017) benchmark, as illustrated in Figure 1(b).

In summary, our contributions are as follows:

- We present RRT-MVS, the first multi-view stereo (MVS) method to use Transformers for cost volume regularization separately in both depth and spatial dimensions.
- We propose a novel Recurrent Regularization Transformer (RRT), which effectively addresses the issue of depth discontinuities in cascade structures of Transformer-based cost volume regularization methods.
- RRT can serve as a plug-in to enhance the performance of cascade-based MVS methods.
- RRT-MVS achieves state-of-the-art performance on both the DTU dataset and the Tanks-and-Temples benchmark.

Related Works

Learning-based MVS

With the rapid development of deep learning techniques, learning-based Multi-View Stereo (MVS) methods have shown more promising results. MVSNet (Yao et al. 2018) introduced an end-to-end deep learning architecture for modeling MVS tasks. However, it requires the use of 3D CNN for cost volume regularization, which demands a large amount of GPU memory, limiting its application in high-resolution scenarios. To address this issue, alternative MVS structures have been developed. On one hand, RNN-based MVS methods (Yao et al. 2019; Wang et al. 2022a; Wang, Li, and Dai 2022; Yan et al. 2023; Cai et al. 2023) use GRU and 2D CNNs iteratively regularize the cost volume, reducing GPU memory consumption but increasing computational complexity. On the other hand, cascade-based MVS methods (Gu et al. 2020; Cheng et al. 2020; Yang et al. 2020) follow a coarse-to-fine strategy for cost volume regularization, resulting in more accurate depth map predictions while maintaining efficiency. Cascade-based MVS methods have been widely embraced in subsequent variants (Peng et al. 2022; Su and Tao 2023; Ye et al. 2023). Moreover, patchmatch concepts (Wang et al. 2021; Li et al. 2024), visibility information (Xu et al. 2022b; Zhang et al. 2023a), uncertainty estimation (Su, Xu, and Tao 2022), geometry priors (Zhang et al. 2023b; Jiang et al. 2024), innovative depth sampling strategies (Wang et al. 2022b; Mi, Di, and Xu 2022), signed distance function technology (Zhang, Zhu, and Lin 2023), and attention mechanism (Ding et al. 2022) have also been integrated into the MVS field to predict more accurate depth maps.

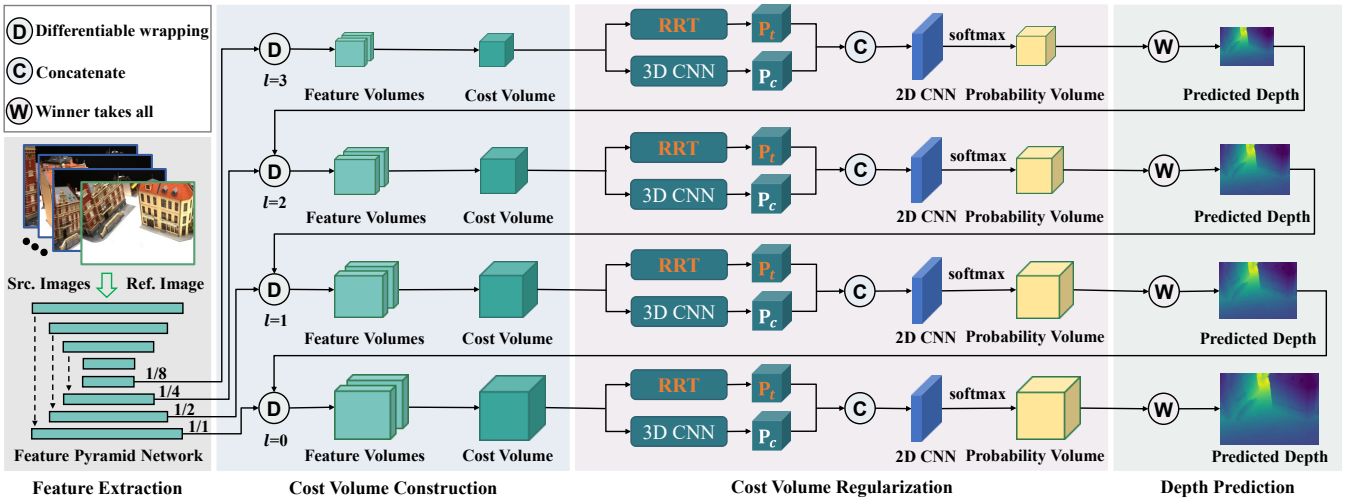

Figure 3: Pipeline of RRT-MVS, which predicted depth maps in a coarse-to-fine manner. RRT-MVS combined the proposed Recurrent Regularization Transformer (RRT) with the standard 3D CNN for cost volume regularization.

Cost Volume Regularization in MVS

Cost volume regularization plays a crucial role in learning-based Multi-View Stereo (MVS) pipelines, which aims to filter redundant matching information in cost volume. Cascade-based MVS methods often employ 3D CNNs for cost volume regularization. To enhance regularization performance, recent approaches such as PatchmatchNet (Wang et al. 2021) have integrated deformable convolutions (Dai et al. 2017) to facilitate spatial matching cost aggregation, while NP-CVP-MVSNet (Yang, Alvarez, and Liu 2022) has utilized sparse convolutions to consolidate matching costs within individual volumes. Despite their advantages, CNN-based methods are limited to processing only local information. GoMVS (Wu et al. 2024) leverages surface normals to address local geometrical inconsistencies, though at a high GPU memory cost. The advent of Transformers (Vaswani et al. 2017), known for their robust global representation capabilities, has brought about a significant shift in the cost volume regularization process. For instance, WT-MVSNet (Liao et al. 2022) proposed a cost Transformer for global feature aggregation, while CostFormer (Chen et al. 2023) aggregate long-range features within the cost volume. MVSFormer++ (Cao, Ren, and Fu 2024) utilizes a pure Transformer architecture for cost volume regularization. Existing Transformer-based cost volume regularization methods handle the depth and spatial dimensions together to achieve global sequence processing through Transformers, as shown in Figure 2. However, the depth dimension does not exhibit the same continuity as the spatial dimension, which is particularly evident at object boundaries and in the background, potentially leading to suboptimal results.

Methodology

Overview

Given N input images $\{\mathbf{I}_i\}_{i=0}^{N-1} \in \mathbb{R}^{3\times H\times W}$, consisting of a reference image and $N-1$ source images, along with the camera intrinsic parameter matrix $\{\mathbf{K}_i\}_{i=0}^{N-1}$, and the associated rotation matrix $R_{0,i}$ and translation vector $t_{0,i}$, where H and W denote the height and width of the input images. The proposed RRT-MVS pipeline is illustrated in Figure 3. Initially, a 4-layer feature pyramid network (Kim et al. 2018) is employed to extract multi-scale feature maps $\{\mathbf{F}_i\}_{i=0}^{N-1} \in \mathbb{R}^{C\times \frac{H}{2^l}\times \frac{W}{2^l}}$ from all input images for depth map estimation in a coarse-to-fine manner, where l denotes the feature layer number and C denotes the feature channels. Subsequently, a cost volume is constructed based on the feature maps, camera parameters, and depth hypotheses $\{\mathbf{d}_i\}_{i=0}^{D-1}$, where D denotes the number of depth hypotheses. The next step involves utilizing the Recurrent Regularization Transformer (RRT) and 3D CNN to regularize the cost volume, resulting in the regularized cost volumes $\mathbf{P}_t$ and $\mathbf{P}_c$. These volumes are then concatenated, and a 2D convolution layer followed by a softmax operation is applied to obtain the final probability volume $\mathbf{P}$. Finally, we use the winner-takes-all strategy to predict the depth map $\mathbf{D}$ from $\mathbf{P}$:

$$\mathbf{D} = \underset{d\in\{d_i\}_{i=0}^{D-1}}{\arg\max}\, \mathbf{P}(d). \tag{1}$$

Cost Volume Construction

To construct the cost volume, we warp the source-view images to multiple hypothesis planes parallel to the reference view based on the number of depth hypotheses. Each depth hypothesis has one hypothesis plane. For each depth hypothesis $\mathbf{d}_i$, we calculate the corresponding pixel $\hat{\mathbf{p}}_i$ on the depth hypothesis plane for the pixel $\mathbf{p}_i$ of the source view.

$$\hat{\mathbf{p}}_i = \mathbf{K}_i \left[\mathbf{R}_{0,i}\left(\mathbf{K}_0^{-1}\mathbf{p}_i\mathbf{d}_i\right) + \mathbf{t}_{0,i}\right], \tag{2}$$

where $\mathbf{K}_i$ denote the intrinsic parameter matrix, $\mathbf{R}_{0,i}$ and $\mathbf{t}_{0,i}$ denote the rotation matrix and translation vector from the ith source view to the reference view. Following previous work (Guo et al. 2019; Xu and Tao 2020a), feature maps are partitioned into $\mathbf{G}$ groups at each feature layer along the channel dimension, assigning $\mathbf{C}/\mathbf{G}$ channels to each group.

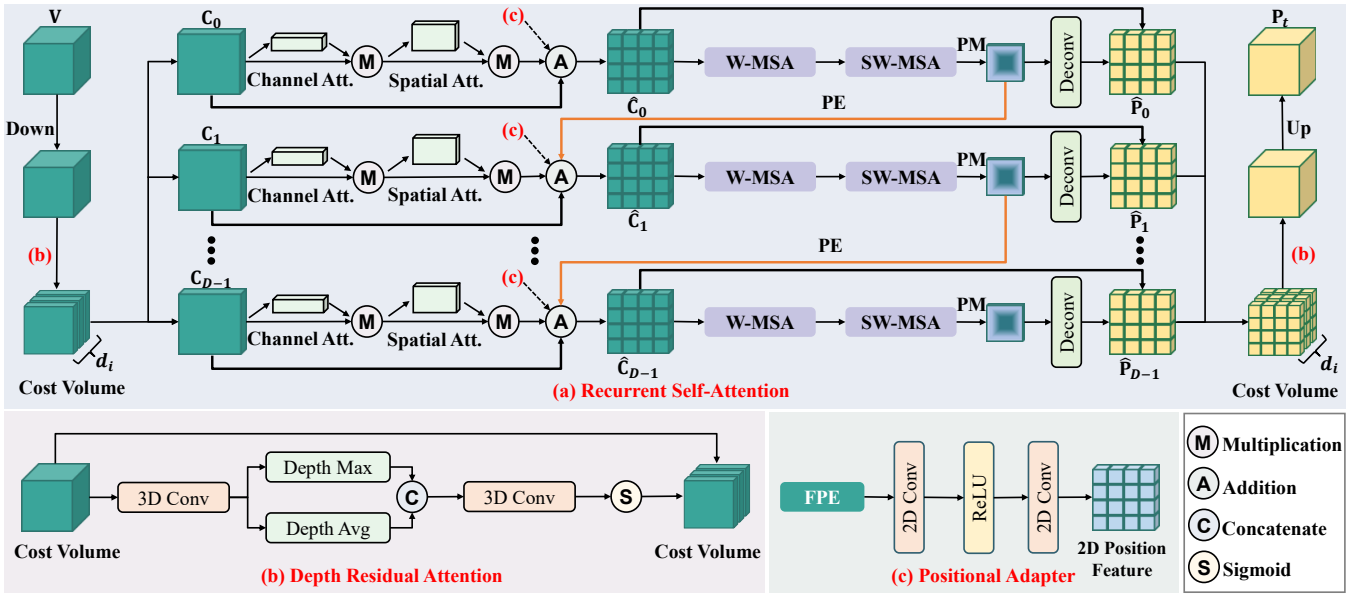

Figure 4: The detailed structure of the proposed Recurrent Regularization Transformer (RRT) consists of three parts: (a) Recurrent Self-Attention (R-SA), (b) Depth Residual Attention (DRA), and (c) Positional Adapter (PA).

The pair-wise group feature correlation between the reference feature volume $\hat{\mathbf{F}}_0^g$ and the ith source feature volume $\hat{\mathbf{F}}_i^g$ is defined as:

$$\mathbf{S}_i^g = \frac{\mathbf{G}}{\mathbf{C}} \left\langle \hat{\mathbf{F}}_i^g, \hat{\mathbf{F}}_0^g \right\rangle, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. We then calculate parameter-free attention weights $\{\mathbf{W}_i\}_{i=1}^{N-1}$ (Wang et al. 2022c) for each correlation. The final cost volume $\mathbf{V} \in \mathbb{R}^{G \times D \times \frac{H}{2^t} \times \frac{W}{2^t}}$ is then constructed as follows:

$$\mathbf{V} = \frac{\sum_{i=1}^{N-1} \mathbf{W}_i \mathbf{S}_i^g}{\sum_{i=1}^{N-1} \mathbf{W}_i}. \tag{4}$$

Recurrent Regularization Transformer (RRT)

The constructed cost volume is inevitably affected by incorrect matches, resulting in noisy estimated depth maps. A typical approach is to use 3D CNNs for regularization to filter out the noise. However, CNN-based regularization methods are limited by the size of their receptive fields, which prevents them from globally perceiving the cost volume. To address this, we propose the Recurrent Regularization Transformer to perform global regularization along the depth dimension, as shown in Figure 4. In practice, we use a non-overlapping patch-wise convolution layer to downsample the cost volume at each stage to the same shape of (M, D, $\frac{H}{8}$, $\frac{W}{8}$), where M denotes intermediate feature channels. Notably, we maintain depth dimension to keep each depth hypothesis independent. After that, we propose the Recurrent Self-Attention (R-SA) to regularize the cost map at each depth hypothesis plane, thereby aggregating cost matching and filtering out redundant feature correlations caused by incorrect matches. Then, we enhance connections between cost maps by applying Depth Residual Attention (DRA) before and after performing R-SA to aggregate depth correlations. Additionally, we introduce a Position Adapter (PA) to improve the 3D positional awareness of 2D cost maps.

Recurrent Self-Attention (R-SA) As shown in Figure 4 (a), we first decompose the cost volume along the depth dimension into cost maps $\{\mathbf{C}_i\}_{i=0}^{D-1} \in \mathbb{R}^{M \times \frac{H}{8} \times \frac{W}{8}}$, where each cost map corresponds to a complete depth hypothesis $\{\mathbf{d}_i\}_{i=0}^{D-1}$. Then, we aggregate the matching information within each cost map using channel attention and spatial attention (Woo et al. 2018), and enhance the feature representation through skip connections to obtain the enhanced cost maps $\widehat{\mathbf{C}}_i$. We then apply window-based multi-head self-attention (W-MSA) and shifted window-based multi-head self-attention (SW-MSA) (Liu et al. 2021) for global regularization of $\widehat{\mathbf{C}}_i$ to obtain regularized cost map $\widehat{\mathbf{P}}_i$. The cost map regularization process for the initial depth hypothesis $\mathbf{d}_0$ and subsequent hypotheses $\mathbf{d}_i$ is formulated as follows:

$$\widehat{\mathbf{P}}_0 = \mathrm{PM}((\text{SW-MSA}(\text{W-MSA}(\widehat{\mathbf{C}}_0))), \tag{5}$$

$$\widehat{\mathbf{P}}_i = \mathrm{PM}(\text{SW-MSA}(\text{W-MSA}(\widehat{\mathbf{C}}_i + \mathrm{PE}(\widehat{\mathbf{P}}_{i-1}))), \tag{6}$$

where PM denotes patch merging for downsampling, PE means patch expanding for upsampling. For each regularized cost map $\widehat{\mathbf{P}}_i$, we perform deconvolutions and apply skip connections to restore the resolution. Then, we aggregate all the $\widehat{\mathbf{P}}_i$ into a regularized cost volume $\mathbf{P}_t$ along the depth dimension, and use a non-overlapping patch-wise transposed convolution layer to upsample the $\mathbf{P}_t$ to the original size.

Depth Residual Attention (DRA) As illustrated in Figure 4(b), we designed a Depth Residual Attention (DRA) module to maximizes the correlation between different depth hy-

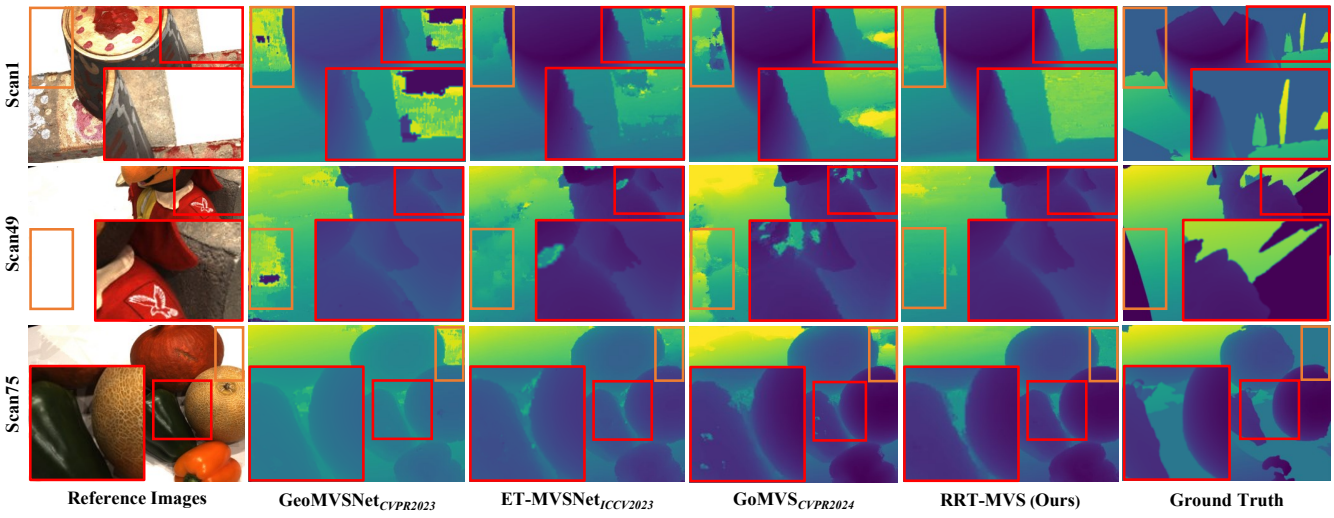

Figure 5: Qualitative comparison of predicted depth maps with state-of-the-art methods (Zhang et al. 2023b; Liu et al. 2023; Wu et al. 2024) on scan1, scan49, and scan75 in the DTU test set. Our method predicts precise boundaries and clear background.

potheses. Initially, we extract cost matching information using a 3D convolution layer, then compute the maximum and average values along the depth dimension and concatenate them. Subsequently, a spatial convolution layer generates the attention map, which is normalized using a sigmoid function. Finally, we apply residual connections to obtain the final cost volume.

Positional Adapter (PA) To enhance the 3D perception capability of 2D cost maps, we employ a lightweight positional adapter to encode 3D information into each cost map, as illustrated in Figure 4(c). First, we calculate the Frustoconical Positional Encoding (FPE) (Cao, Ren, and Fu 2024) based on the cost volume at each stage. FPE contains the 3D positional information of the scene. We downsample and reshape FPE to (CD, $\frac{H}{8}$, $\frac{W}{8}$) to match the size of the 2D cost maps. Then, we apply two simple 2D convolutions and ReLU activation to construct the 2D positional feature, with a shape of (M, $\frac{H}{8}$, $\frac{W}{8}$). Finally, we add the 2D positional features to the regularization process of each cost map to enhance the 3D positional awareness of the proposed R-SA.

Loss Function

We use cross entropy loss to supervise the probability volume for all stages.

$$Loss = \sum_{\mathbf{q}\in\{\mathbf{q}_v\}} \sum_{i=0}^{D-1} -G^{(d_i)}(\mathbf{q}) \log \left[P^{(d_i)}(\mathbf{q}) \right], \quad (7)$$

where $G^{(d_i)}(\mathbf{q})$ and $P^{(d_i)}(\mathbf{q})$ denotes ground truth volume and probability volume of depth hypotheses d_i at pixel q. $\{\mathbf{q}_v\}$ denotes a set of valid ground truth pixels.

Experiments

Datasets and Metrics

Datasets DTU (Aanæs et al. 2016) dataset includes 128 indoor scenes, each captured from 49 or 64 viewpoints under 7 different lighting conditions using a fixed camera trajectory. As described in (Yao et al. 2018), the dataset is divided into training, test, and validation sets, with a total of 27,097 training samples. Tanks-and-Temples (Knapitsch et al. 2017) dataset is an extensive collection of real-world scenes divided into an intermediate subset and an advanced subset. BlendedMVS (Yao et al. 2020) dataset is a large-scale synthetic dataset with both indoor and outdoor scenes, consisting of training and validation data.

Metrics We evaluate the accuracy (Acc.) and completeness (Comp.) of the generated point clouds based on distance metrics on the DTU dataset (Aanæs et al. 2016). The overall metric refers to the average of the aforementioned two. We also report the Mean Absolute Error (MAE) and depth error ratios with distance thresholds of 2mm (e_2), 4mm (e_4), and 8mm (e_8) to measure the precision of depth maps. For the Tanks-and-Temples (Knapitsch et al. 2017) benchmark, we report F-score represented in percentages.

Implementation Details

Training Our network was developed using PyTorch (Paszke et al. 2019) and employed the Adam (Kingma and Ba 2014) optimizer. Following standard procedures, we initially trained our model on the DTU (Aanæs et al. 2016) training set using 5-view images with a resolution of 512×640. We started with a learning rate of 0.001 for 10 epochs with a batch size of 2. Subsequently, we fine-tuned the model on the BlendedMVS (Yao et al. 2020) dataset using 11-view images with a resolution of 576×768. This phase began with a learning rate of 0.001 for an additional 15 epochs with a batch size of 2. Our approach involved inverse depth sampling ranging from $425mm$ to $935mm$, with 4 coarse-to-fine stages that incorporated both depth hypotheses and group feature correlations of 8-8-4-4.

Evaluation We trained our model on the DTU training set and evaluated its performance on the DTU test set using 5-

Methods	Acc.↓	Comp.↓	**Overall↓**
Gipuma$_{ICCV2015}$	**0.283**	0.873	0.578
COLMAP$_{CVPR2016}$	0.400	0.664	0.532
UniMVSNet$_{CVPR2022}$	0.352	0.278	0.315
TransMVSNet$_{CVPR2022}$	0.321	0.289	0.305
MVSTER$_{ECCV2022}$	0.350	0.276	0.313
DispMVS$_{AAAI2023}$	0.354	0.324	0.339
EPNet$_{AAAI2023}$	0.299	0.323	0.311
CostFormer$_{IJCAI2023}$	0.301	0.322	0.312
RA-MVSNet$_{CVPR2023}$	0.326	0.268	0.297
GeoMVSNet$_{CVPR2023}$	0.331	0.259	0.295
DMVSNet$_{ICCV2023}$	0.338	0.272	0.305
ET-MVSNet$_{ICCV2023}$	0.329	0.253	0.291
MVSFormer$_{TMLR2023}$	0.327	0.251	0.289
DS-PMNet$_{AAAI2024}$	0.323	0.257	0.290
GoMVS$_{CVPR2024}$	0.347	**0.227**	0.287
RRT-MVS (Ours)	0.309	0.261	**0.285**

Table 1: Quantitative results of reconstructed point clouds on the DTU test set with distance metrics [mm].

Methods	Overall↓	Mem.(MB)↓	Mem.(%)↓
WT-MVSNet	-0.013	+1204	+29.97
CostFormer	-0.009	+370	+15.93
MVSFormer++	-0.003	+994	+20.00
GoMVS	**-0.028**	+9114	+278.13
RRT-MVS (Ours)	**-0.028**	**+266**	**+7.98**

Table 2: Quantitative improvement of overall metric [mm] and additional GPU memory (Mem.) for different cost volume regularization methods on the DTU test set.

view images with a resolution of 832×1152, which required 0.35s and 3.5GB of memory on a NVIDIA RTX 3090 GPU. Subsequently, we fine-tuned the model on the BlendedMVS dataset and assessed its generalization ability on the Tanks-and-Temples (Knapitsch et al. 2017) benchmark.

Results on DTU

To assess the effectiveness of our method, we employed the dynamic check strategy (Yan et al. 2020) for depth fusion on the DTU (Aanæs et al. 2016) test set. As shown in Table 1, RRT-MVS achieved the highest overall score. As illustrated in Figure 5, RRT-MVS predicts clear boundaries (red bounding box) and backgrounds with less noise (orange bounding box). This indicates that our method can effectively perform global cost volume regularization for regions with depth discontinuities. The qualitative point clouds comparison is shown in Figure 6.

Comparison with Other Regularization Methods We also compare RRT with cost volume regularization methods proposed in other MVS methods, including Transformer-based methods: WT-MVSNet (Liao et al. 2022), CostFormer (Chen et al. 2023), MVSFormer++ (Cao, Ren, and Fu 2024), and the CNN-based method GoMVS (Wu et al. 2024). For a fair comparison, we focus solely on the performance enhancements and additional GPU memory consumption attributed to the cost volume regularization methods as outlined in the original papers. As shown in Table 2, RRT-MVS achieves at least a 2× improvement in overall performance compared to existing Transformer-based cost volume regularization methods, while reducing GPU memory consumption. Furthermore, compared to the current best CNN-based method, GoMVS(Wu et al. 2024), RRT-MVS reduces GPU memory consumption by an impressive 97%.

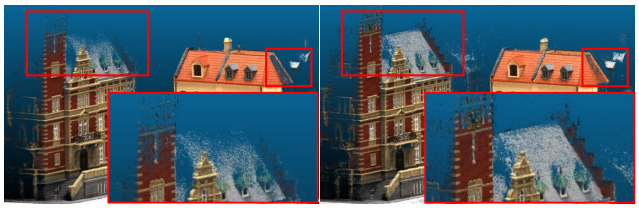

Figure 6: Qualitative comparison with the state-of-the-art method (Wu et al. 2024) of reconstructed point clouds on scan 29 from the DTU test set. Our method reconstructs more accurate and complete details in the boundary regions.

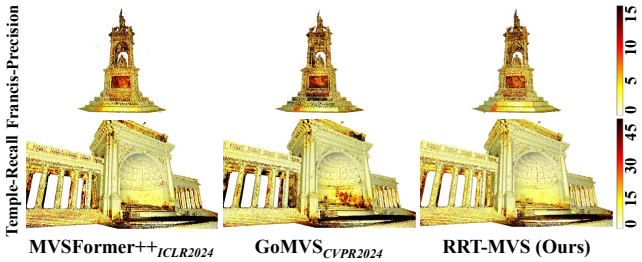

Figure 7: Qualitative comparisons of point clouds with top-2 published methods (Cao, Ren, and Fu 2024; Wu et al. 2024) on the Tanks-and-Temples benchmark. Brighter areas indicate smaller errors with distance threshold [mm].

Results on Tanks-and-Temples

To evaluate the generalization capability of our method in complex environments, we employed a standard dynamic check strategy for depth fusion on the Tanks-and-Temples benchmark (Knapitsch et al. 2017). As presented in Table 3, our method achieved the highest mean F-score of **68.16** and **43.29** on the intermediate and the advanced subsets, respectively. Notably, the Tanks-and-Temples dataset includes outdoor scenes with intricate backgrounds, demonstrating the effectiveness of our method in managing depth discontinuities. Furthermore, as illustrated in Figure 7, RRT-MVS exhibits superior precision and recall compared to existing state-of-the-art methods.

Ablation Study

We conducted quantitative ablation experiments to validate the effectiveness of the RRT-MVS design. All ablation experiments were performed on the DTU (Aanæs et al. 2016) dataset with same parameters, reconstructing point clouds using the normal fusion strategy (Schönberger et al. 2016).

Methods	Intermediate Subset↑									Advanced Subset↑						
	Mean	Fam.	Fran.	Hor.	Lig.	M60	Pan.	Pla.	Tra.	**Mean**	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
COLMAP$_{CVPR2016}$	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04	27.24	16.02	25.23	34.70	41.51	18.05	27.94
ACMP$_{AAAI2020}$	58.41	70.30	54.06	54.11	61.65	54.16	57.60	58.12	57.25	37.44	30.12	34.68	**44.58**	50.64	27.20	37.43
TransMVSNet$_{CVPR2022}$	63.52	80.92	65.83	56.94	62.54	63.06	60.00	60.20	58.67	37.00	24.84	44.59	34.77	46.49	34.69	36.62
UniMVSNet$_{CVPR2022}$	64.36	81.20	66.43	53.11	63.46	66.09	64.84	62.23	57.53	38.96	28.33	44.36	39.74	52.89	33.80	34.63
MVSTER$_{ECCV2022}$	60.92	80.21	63.51	52.30	61.38	61.47	58.16	58.98	51.38	37.53	26.68	42.14	35.65	49.37	32.16	39.19
DispMVS$_{AAAI2023}$	59.07	74.73	60.67	54.13	59.58	58.02	53.39	58.63	53.42	34.90	26.09	38.01	33.19	44.90	28.49	38.75
EPNet$_{AAAI2023}$	63.68	77.78	59.61	58.87	66.17	65.58	63.53	60.34	57.56	40.52	30.24	45.52	40.83	53.51	32.57	40.43
CostFormer$_{IJCAI2023}$	64.51	81.31	65.65	55.57	63.46	66.24	65.39	61.27	57.30	39.43	29.18	45.21	39.88	53.38	34.07	34.87
RA-MVSNet$_{CVPR2023}$	65.72	82.44	66.61	58.40	64.78	**67.14**	65.60	62.74	58.08	39.93	29.14	46.04	40.30	53.22	34.63	36.28
GeoMVSNet$_{CVPR2023}$	65.89	81.64	67.53	55.78	68.02	65.49	67.19	63.27	58.22	41.52	30.23	46.54	39.98	53.05	35.98	43.34
DMVSNet$_{ICCV2023}$	64.66	81.27	67.54	59.10	63.12	64.64	64.80	59.83	56.97	41.17	30.08	46.10	40.65	53.53	35.08	41.60
ET-MVSNet$_{ICCV2023}$	65.49	81.65	68.79	59.46	65.72	64.22	64.03	61.23	58.79	40.41	28.86	45.18	38.66	51.10	35.39	43.23
SD-MVS$_{AAAI2024}$	63.31	75.37	63.37	61.53	65.71	62.30	62.37	59.27	56.58	40.18	31.17	44.26	39.41	51.96	32.66	41.64
DS-PMNet$_{AAAI2024}$	64.16	81.11	63.43	60.84	62.23	64.96	61.92	61.41	57.35	39.78	28.52	44.93	39.12	51.68	33.77	40.67
MVSFormer++$_{ICLR2024}$	67.18	**82.69**	69.44	64.24	69.16	64.13	66.43	61.19	60.12	41.60	29.93	45.69	39.46	53.58	35.56	45.39
GoMVS$_{CVPR2024}$	66.44	82.68	69.23	**69.19**	63.56	65.13	62.10	58.81	**60.80**	43.07	**35.52**	**47.15**	42.52	52.08	36.34	44.82
RRT-MVS (Ours)	**68.16**	82.54	**72.31**	61.44	**69.89**	65.32	**68.88**	**64.45**	60.48	**43.29**	30.95	46.42	41.13	**55.46**	**37.63**	**48.12**

Table 3: Quantitative results on the Tanks-and-Temples benchmark with F-score [%]. Bold figures represent the best and underline figures represent the second best, respectively. The mean refers the average F-score of all scenes.

Settings			Acc.↓	Comp.↓	Overall↓	MAE↓
R-SA	DRA	PA				
			0.351	0.284	0.318	6.64
✓			0.325	0.272	0.299	6.35
✓	✓		0.319	0.284	0.302	5.71
✓		✓	**0.320**	0.270	0.295	5.60
✓	✓	✓	**0.320**	**0.260**	**0.290**	**5.24**

Table 4: Ablation study of each component on RRT.

Modules	Overall↓	MAE↓	e_2↓	e_4↓	e_8↓
3D CNN	0.318	6.64	20.60	13.36	9.00
RRT	0.311	5.28	16.29	**9.36**	**6.05**
Combination	**0.290**	**5.24**	**16.12**	9.52	6.21

Table 5: Ablation study of combining RRT and 3D CNN.

Effectiveness of each component in RRT We conducted experiment on RRT to verify the effectiveness of each component. As detailed in Table 4, incorporating R-SA significantly improved model performance. While DRA alone enhanced the depth map quality, the addition of PA further refined the point cloud quality. This highlights the complementary roles of DRA and PA in enhancing depth and spatial dimensions. Ultimately, the integration of all three components resulted in state-of-the-art performance.

Effectiveness of combining RRT with 3D CNN We also examined the necessity of integrating RRT with 3D CNNs. As shown in Table 5, RRT significantly improved the depth map quality compared to using the 3D CNN alone, although the enhancement in point cloud quality was less pronounced. This limitation is due to RRT's lack of local detail understanding. Ultimately, the combined RRT-MVS, which integrates RRT with 3D CNN, demonstrated improvements in both depth map and point cloud quality.

Methods	Acc.↓	Comp.↓	Overall↓
CasMVSNet	0.325	0.385	0.355
CasMVSNet+RRT	0.346	0.306	0.326
MVSTER	0.350	0.276	0.313
MVSTER+RRT	0.352	0.250	0.301

Table 6: Compatibility study of RRT as a plug-in.

Effectiveness of RRT as a Plug-in The proposed Recurrent Regularization Transformer (RRT) can be used as a plug-in to enhance cascade-based MVS methods. We validated this by incorporating RRT into the classic three-stage CasMVSNet (Gu et al. 2020) and the four-stage MVSTER (Wang et al. 2022c) backbone networks, ensuring consistency with their original settings. RRT was combined with the original 3D CNNs for cost volume regularization. The experimental results, presented in Table 6, indicate performance improvements in both architectures, demonstrating the effectiveness of RRT as a plug-in module.

Conclusion

In this paper, we present a novel Recurrent Regularization Transformer (RRT) for global cost volume regularization in Multi-View Stereo (MVS). Specifically, we introduce recurrent self-attention for cost map regularization, enhance depth correlations through depth residual attention, and utilize a positional adapter to improve the 3D positional awareness of 2D cost maps. We integrate RRT into the baseline network to construct RRT-MVS. Experimental results demonstrate that RRT-MVS achieves state-of-the-art performance on multiple MVS datasets. Additionally, RRT can serve as a plug-in for MVS networks, significantly outperforming other cost volume regularization methods.

Acknowledgments

This work was supported by the National Nature Science Foundation of China under Grant 62227801.

References

Aanæs, H.; Jensen, R. R.; Vogiatzis, G.; Tola, E.; and Dahl, A. B. 2016. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120: 153–168.

Cai, C.; Ji, P.; Yan, Q.; and Xu, Y. 2023. Riav-mvs: Recurrent-indexing an asymmetric volume for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 919–928.

Cao, C.; Ren, X.; and Fu, Y. 2022. MVSFormer: Multi-View Stereo by Learning Robust Image Features and Temperature-based Depth. *Transactions on Machine Learning Research*.

Cao, C.; Ren, X.; and Fu, Y. 2024. MVSFormer++: Revealing the Devil in Transformer's Details for Multi-View Stereo. *arXiv preprint arXiv:2401.11673*.

Chen, W.; Xu, H.; Zhou, Z.; Liu, Y.; Sun, B.; Kang, W.; and Xie, X. 2023. CostFormer: cost transformer for cost aggregation in multi-view stereo. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 599–608.

Cheng, S.; Xu, Z.; Zhu, S.; Li, Z.; Li, L. E.; Ramamoorthi, R.; and Su, H. 2020. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2524–2534.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 764–773.

Dao, T. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Ding, Y.; Yuan, W.; Zhu, Q.; Zhang, H.; Liu, X.; Wang, Y.; and Liu, X. 2022. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8585–8594.

Galliani, S.; Lasinger, K.; and Schindler, K. 2015. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, 873–881.

Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; and Tan, P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2495–2504.

Guo, X.; Yang, K.; Yang, W.; Wang, X.; and Li, H. 2019. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3273–3282.

Jiang, J.; Cao, M.; Yi, J.; and Li, C. 2024. DI-MVS: Learning Efficient Multi-View Stereo With Depth-Aware Iterations. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3180–3184.

Kim, S.-W.; Kook, H.-K.; Sun, J.-Y.; Kang, M.-C.; and Ko, S.-J. 2018. Parallel Feature Pyramid Network for Object Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 234–250.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13.

Li, H.; Guo, Y.; Zheng, X.; and Xiong, H. 2024. Learning Deformable Hypothesis Sampling for Accurate PatchMatch Multi-View Stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3082–3090.

Liao, J.; Ding, Y.; Shavit, Y.; Huang, D.; Ren, S.; Guo, J.; Feng, W.; and Zhang, K. 2022. Wt-mvsnet: window-based transformers for multi-view stereo. *Advances in Neural Information Processing Systems*, 35: 8564–8576.

Liu, T.; Ye, X.; Zhao, W.; Pan, Z.; Shi, M.; and Cao, Z. 2023. When Epipolar Constraint Meets Non-local Operators in Multi-View Stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18088–18097.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Mi, Z.; Di, C.; and Xu, D. 2022. Generalized binary search network for highly-efficient multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12991–13000.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Peng, R.; Wang, R.; Wang, Z.; Lai, Y.; and Wang, R. 2022. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8645–8654.

Schönberger, J. L.; Zheng, E.; Frahm, J.-M.; and Pollefeys, M. 2016. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 501–518. Springer.

Su, W.; and Tao, W. 2023. Efficient edge-preserving multi-view stereo network for depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2348–2356.

Su, W.; Xu, Q.; and Tao, W. 2022. Uncertainty guided multi-view stereo network for depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7796–7808.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, F.; Galliani, S.; Vogel, C.; and Pollefeys, M. 2022a. IterMVS: Iterative probability estimation for efficient multi-view stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8606–8615.

Wang, F.; Galliani, S.; Vogel, C.; Speciale, P.; and Pollefeys, M. 2021. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14194–14203.

Wang, L.; Gong, Y.; Ma, X.; Wang, Q.; Zhou, K.; and Chen, L. 2022b. Is-mvsnet: importance sampling-based mvsnet. In *European Conference on Computer Vision*, 668–683. Springer.

Wang, S.; Li, B.; and Dai, Y. 2022. Efficient multi-view stereo by iterative dynamic cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8655–8664.

Wang, X.; Zhu, Z.; Huang, G.; Qin, F.; Ye, Y.; He, Y.; Chi, X.; and Wang, X. 2022c. MVSTER: Epipolar transformer for efficient multi-view stereo. In *European Conference on Computer Vision*, 573–591. Springer.

Wang, Y.; Zeng, Z.; Guan, T.; Yang, W.; Chen, Z.; Liu, W.; Xu, L.; and Luo, Y. 2023. Adaptive patch deformation for textureless-resilient multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1621–1630.

Wei, Z.; Zhu, Q.; Min, C.; Chen, Y.; and Wang, G. 2021. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6187–6196.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, J.; Li, R.; Xu, H.; Zhao, W.; Zhu, Y.; Sun, J.; and Zhang, Y. 2024. GoMVS: Geometrically Consistent Cost Aggregation for Multi-View Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20207–20216.

Xu, Q.; Kong, W.; Tao, W.; and Pollefeys, M. 2022a. Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4945–4963.

Xu, Q.; Su, W.; Qi, Y.; Tao, W.; and Pollefeys, M. 2022b. Learning Inverse Depth Regression for Pixelwise Visibility-Aware Multi-View Stereo Networks. *International Journal of Computer Vision*, 130(8): 2040–2059.

Xu, Q.; and Tao, W. 2020a. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12508–12515.

Xu, Q.; and Tao, W. 2020b. Planar prior assisted patchmatch multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12516–12523.

Yan, J.; Wei, Z.; Yi, H.; Ding, M.; Zhang, R.; Chen, Y.; Wang, G.; and Tai, Y.-W. 2020. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European conference on computer vision*, 674–689. Springer.

Yan, Q.; Wang, Q.; Zhao, K.; Li, B.; Chu, X.; and Deng, F. 2023. Rethinking disparity: a depth range free multi-view stereo based on disparity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3091–3099.

Yang, J.; Alvarez, J. M.; and Liu, M. 2022. Non-parametric depth distribution modelling based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8626–8634.

Yang, J.; Mao, W.; Alvarez, J. M.; and Liu, M. 2020. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4877–4886.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, 767–783.

Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; and Quan, L. 2019. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5525–5534.

Yao, Y.; Luo, Z.; Li, S.; Zhang, J.; Ren, Y.; Zhou, L.; Fang, T.; and Quan, L. 2020. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1790–1799.

Ye, X.; Zhao, W.; Liu, T.; Huang, Z.; Cao, Z.; and Li, X. 2023. Constraining Depth Map Geometry for Multi-View Stereo: A Dual-Depth Approach with Saddle-shaped Depth Cells. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17661–17670.

Yuan, Z.; Cao, J.; Li, Z.; Jiang, H.; and Wang, Z. 2024. SD-MVS: Segmentation-Driven Deformation Multi-View Stereo with Spherical Refinement and EM Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6871–6880.

Zhang, J.; Li, S.; Luo, Z.; Fang, T.; and Yao, Y. 2023a. Vis-mvsnet: Visibility-aware multi-view stereo network. *International Journal of Computer Vision*, 131(1): 199–214.

Zhang, Y.; Zhu, J.; and Lin, L. 2023. Multi-View Stereo Representation Revist: Region-Aware MVSNet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17376–17385.

Zhang, Z.; Peng, R.; Hu, Y.; and Wang, R. 2023b. GeoMVSNet: Learning Multi-View Stereo With Geometry Perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21508–21518.