Long-range Attention Network for Multi-View Stereo

Xudong Zhang[1] Yutao Hu[1] Haochen Wang[1] Xianbin Cao[1,2,3*]
Baochang Zhang[4]
[1]School of Electronic and Information Engineering, Beihang University, Beijing, China
[2]Key Laboratory of Advanced Technology of Near Space Information System,
Ministry of Industry and Information Technology of China
[3]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, China
[4]Beihang University, Beijing, China
{xdzhang,huyutao,haochenwang,xbcao,bczhang}@buaa.edu.cn

Abstract

Learning-based multi-view stereo (MVS) has recently gained great popularity, which can efficiently infer depth map and reconstruct fine-grained scene geometry. Previous methods calculate the variance of the corresponding pixel pairs to determine whether they are matched mostly based on the pixel-wise measure, which fails to consider the inter-dependence among pixels and is ineffective on the matching of texture-less or occluded regions. These false matching problems challenge MVS and result in its most failure cases. To address the issues, we introduce a Long-range Attention Network (LANet) to selectively aggregate reference features to each position to capture the long-range interdependence across the entire space. As a result, similar features relate to each other regardless of their distance, propagating more guiding information for the effective match. Furthermore, we introduce a new loss to supervise the intermediate probability volume by constraining its distribution reasonably centered at the true depth. Extensive experiments on large-scale DTU dataset demonstrate that the proposed LANet achieves the new state-of-the-art performance, outperforming previous methods by a large margin. Our method is generic and also achieves comparable results on outdoor Tanks and Temples dataset without any fine-tuning, which validates our method's generalization ability.

1. Introduction

Multi-view stereo (MVS) has recently received growing interest in the computer vision and graphics community for its applications in 3D visualization, autonomous driving and augmented reality etc. MVS aims to reconstruct a 3D geometric representation from a collection of multi-view images with known camera parameters. Recent success of deep learning has inspired researchers to exploit learning-based MVS methods. Most of the current works [34, 22, 32] employ a two-step architecture, i.e., depth map estimation and fusion to form the point cloud.

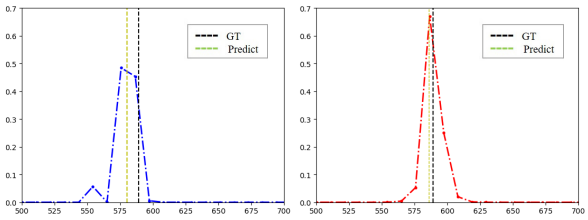

Figure 1. Visual comparison of predicted probability volume distribution at one pixel between CasMVSNet [14] and ours. The left picture is the output of CasMVSNet, which is scattered and cannot be centralized to the ground truth, while the right picture is the output of our LANet reasonably centered at the true depth.

The core of depth map estimation is to construct a cost volume to measure the similarity between corresponding image patches. The basic idea is to first warp the extracted 2D features of each view to the discrete hypothesis depth planes upon the reference camera frustum using the plane-sweep algorithm [11], and then aggregate these multiple volumes to one cost volume by the variance-based metric [34, 7, 14]. However, this pixel-wise operation ignores the interdependence among pixels, since its calculation of matching degree for each pixel pair is produced independently from the other pixel pairs. As a result, the constructed cost volume computed from image features could be noise-contaminated and inconsistent, which is often not conducive to the dense matching on occluded or texture-less regions.

*Corresponding author.

Figure 2. Reconstruction comparison especially in texture-less area between state-of-the-art approaches [35, 14] and our LANet. As shown our method is able to obtain more completeness in texture-less surface (e.g., wall in this example).

Furthermore, previous methods [33, 32, 9] predict depth maps from the *expectation* of probability volume distributions which express the possibility of different depth hypotheses. However, existing volume learning methods are driven by the depth regression, which are implicit and not very efficient, considering that different probability distributions might share the same *expectation*. Also, the optimal distribution should be a Gaussian distribution, which has the highest probability at the true depth. As illustrated in Fig. 1, for those falsely matched pixels, their probability distributions are far from optimal, since they might be not effectively supervised by the intermediate layer. Therefore, it becomes essential to explicitly supervise the probability volume distribution.

In this paper, to deal with the two aforementioned problems, we propose the Long-range Attention Network (LANet) for MVS. We introduce the LANet to capture long-range interdependence between reference and other source images across the entire space efficiently. As shown in Fig. 4, we first gather global descriptors from the reference image and then obtain the attentive features by adaptively aggregating descriptors to each location of other input images through attention vectors. After that, we warp the attentive features to construct plane sweep volumes and then fuse them together to build the 3D cost volume. In this way, pixels in all input images are fully connected with the reference image and the constructed cost volume would be smoother. This provides more guiding information for prediction in the texture-less area. For example, as shown in Fig. 2, our proposed LANet achieves a better completeness in the wall regions. Besides, we employ a 3D convolution network with Atrous Spatial Pyramid Pooling (ASPP) [6, 5] to regularize the cost volume instead of the previous U-Net, incorporating contextual information for depth inference. ASPP module enables the network to effectively enlarge the receptive fields to incorporate long-range context and alleviate object boundaries missing due to the pooling or convolutions with striding operations used in 3D U-Net.

Moreover, we introduce a new loss to supervise the intermediate probability based on a Gaussian distribution, whose *expectation* is the true depth as shown in Fig. 1. Specifically, we apply the score map of the estimation confidence to adjust the sharpness (variance) of the probability distribution. For example, those matched pixels should have a sharp peak (lower variance), while those uncertain ones have a relatively flat peak (higher variance). In this way, unreasonable probability distributions are removed, which can thus enhance the robustness and generalization of our method for new objects during inference.

We conduct extensive experiments on DTU dataset [1] and Tanks & Temples dataset [18] for performance evaluation. The proposed LANet largely surpasses previous state-of-the-art methods. Moreover, we perform thorough ablation studies to demonstrate the benefits of the introduced long-range attention module and the probability volume loss, offering more insight into its great effectiveness for MVS.

The major contributions of this work are summarized as follows:

- We introduce a long-range attention network for MVS to selectively aggregate reference features to each position to capture the interdependence among pixels. Our method instinctively facilitates propagating more guiding information to measure the similarity between reference and other source images.

- We introduce a new loss to supervise the distribution of probability volumes reasonably centered at the true depth, which can enhance robustness of the network to reconstruct new objects.

- We achieve the best overall performance on the DTU and Tanks & Temples benchmarks, largely surpassing the state-of-the-art methods by up to 3.7% for the *overall scores*.

2. Related work

Learning-based MVS. Inspired by the success of deep convolutional neural networks (CNN), several learning-based MVS methods [2, 23, 24, 37, 19] have been presented and have achieved very promising results. The first learning-based MVS pipeline is introduced by [28], which pre-computes the cost volume to learn the probability of voxels lying on the surface. Concurrently, LSM [16] infers the surface voxels by presenting a learnable system to up-project pixel features to the 3D volume and using 3D CNN to regularize. However, they are restricted to only small-scale reconstructions as the volumetric representations lead to expensive memory usage. To solve this problem, depth-map based learning method has been approached by [34]. The authors build a single 3D cost volume with uniformly sampled depth hypotheses by projecting 2D image features into 3D space and then use a stack of 3D CNNs to infer the final depth. Nonetheless, a single cost volume often requires a large number of depth planes to achieve enough reconstruction accuracy, and it is difficult to reconstruct high-resolution depth, limited by the memory bottleneck. Therefore, [14, 33, 9] utilize multi-stage small volumes to progressively regress a high-quality depth map in a coarse-to-fine fashion. The later stage uses the estimated depth map from the earlier stage to adaptively adjust the sampling range of depth hypotheses and construct new high-resolution cost volumes and further boost accuracy. Similar to the recent coarse-to-fine approaches, we adopt the multi-stage structure in our framework. But different from previous works constructing cost volumes by pixel-wise matching, we propose to capture long-range interdependence among pixels and aggregate more guiding information from the entire space to decide whether these corresponding image patches are matched.

Attention. Attention mechanism [10, 31, 15] has recently been incorporated into deep learning for performance augmentation, which has achieved great success in a large variety of tasks. Self-attention mechanism calculates the response at a position in a sequence by attending to all positions within the same sequence, which has been widely used to assign importance to each type of neighbor and reconstruct feature representations. In particular, the work [30] is the first to propose the self-attention mechanism to draw global dependencies of inputs and applies it in machine translation. Meanwhile, attention modules are increasingly applied in image vision flied. Zhang $et\ al.$ [36] introduces self attention mechanism to learn a better image generator. Recent networks [31, 8] capture dependencies in space-time dimension for videos and images, which can be considered as a generalization of self-attention mechanism. Different from previous works, we introduce the long-range attention model considering the spatial dependencies not only in the per-view feature map but also across multi-view features.

3. Method

3.1. Architecture overview

For a MVS task, our goal is to reconstruct a 3D geometric representation from a collection of multi-view images with known camera parameters. As shown in Fig. 3, the proposed LANet is built upon a course-to-fine architecture to progressively predict the fine-grained depth map, which is now widely used in existing learning-based MVS networks [14, 33, 9]. We introduce two major innovative modules into the network: 1) the long-range attention modules are firstly presented to obtain the attentive pyramid feature maps, capturing the long-range and multi-scale correspondence between reference image and other source images; 2) we introduce a new loss to supervise the intermediate probability volume by reasonably constraining it to the Gaussian distribution.

To be more specific, suppose we are provided with a reference image $\mathbf{I}_1$ and $N-1$ other source images $\{\mathbf{I}_n\}_{n=2}^N$. The LANet regresses a fine-grained depth map $\mathbf{D}$ for $\mathbf{I}_1$ from $\{\mathbf{I}_n\}_{n=1}^N$ in a coarse-to-fine manner. As shown in Fig. 4, we first leverage a Feature Pyramid Network [20] to extract multi-scale deep image features at three resolutions $\{\mathbf{f}_n^l\}_{l=1}^3$. Then, each pair of deep features $\{\mathbf{f}_n^l\}_{n=1}^N$ is fed into the proposed long-range attention module (LAM) to obtain attentive pyramid features. Then as illustrated in Fig. 3, we warp these features to the adaptive depth hypotheses based on the prediction of previous stage to construct plane sweep volumes, which are fused together to build the 3D cost volume. Moreover, we apply 3D convolution network with ASPP to regularize the cost volume to predict the depth probability distribution. The depth map is finally referred from the $expectation$ of the per-pixel distribution. We also utilize a novel probability volume loss as an intermediate supervision, coupled with the depth regression loss to supervise the network training.

3.2. Long-range attention

The long-range attention module (LAM) is designed mainly with two steps. The first step is illustrated at the bottom row in Fig. 4, gathering key features of reference $\mathbf{f}_1^l$ into global descriptors $\mathbf{G}^l$ through attention-based second-order pooling [4, 21], which embeds the entire feature space into an informative compact package. The second step aggregates interdependence by adaptively selecting these informative reference descriptors $\mathbf{G}^l$ to each location of input features $\{\mathbf{f}_n^l\}_{n=1}^N$ and obtains the attentive output $\{\mathbf{Z}_n^l\}_{n=1}^N$, as shown at the top N rows in Fig. 4.

Reference global descriptors. Given a reference feature $\mathbf{f}_1 \in \mathbb{R}^{C\times H\times W}$, we first utilize different convolutional layers to embed it to two new feature maps $\mathbf{A}$ and $\mathbf{B}$, respectively, where $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{C\times H\times W}$. Then we reshape them to $\mathbb{R}^{C\times(HW)}$, where HW is the number of pixels.

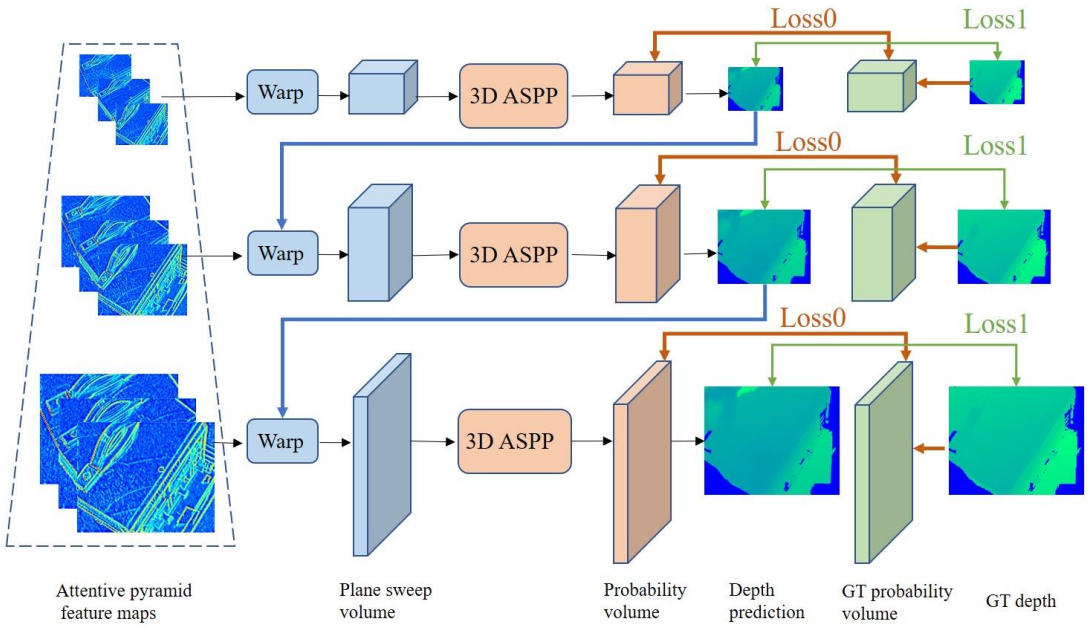

Figure 3. An overview of our proposed LANet. Our LANet utilizes the attentive pyramid features to progressively predict the fine-grained depth maps in a coarse-to-fine manner. The features are obtained from the architecture illustrated in Fig. 4. Specifically, we first warp the features to construct the plane sweep volumes, and then regularize them with 3D ASPP to predict the probability volumes for depth reference. Note the LANet is jointly trained end-to-end by the new novel probability volume loss $Loss0$ and the depth loss $Loss1$.

For ease of explanation, we rewrite $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_{hw}]$ and $\mathbf{B}^\top = [\mathbf{b}_1, ..., \mathbf{b}_c]$, where $\mathbf{a}_i$ is a c-dimension column vector and $\mathbf{b}_j$ is a hw-dimension column vector. The global descriptors $\mathbf{G} = [\mathbf{g}_1, ..., \mathbf{g}_c] \in \mathbb{R}^{C\times C}$ can be calculated by embedding the entire feature space into a compact package through attention-based second-order pooling, where each primitive $\mathbf{g}_j$ is obtained by incorporating local features $\{\mathbf{a}_i\}_{i=1}^{hw}$ weighted by $\mathbf{b}_j$:

$$\mathbf{g}_j = \mathbf{A} \cdot softmax(\mathbf{b}_j) = \sum_{i=1}^{hw} \hat{b}_{j,i} \cdot \mathbf{a}_i, \tag{1}$$

$$\hat{b}_{j,i} = \frac{\exp(b_{j,i})}{\sum_i \exp(b_{j,i})}, \tag{2}$$

where we apply the softmax function to normalize $\mathbf{b}_j$ into the one with unit sum, following second-order attention pooling process and $b_{j,i}$ indicates the i_{th} element of $\mathbf{b}_j$.

This step offers an effective way to capture informative features of reference. For instance, if $\mathbf{b}_j$ is densely attended on all locations, we can acquire the texture and lighting features. On the contrary, if $\mathbf{b}_j$ is sparsely attended on a specific region, we can acquire the semantic features, e.g. an object.

Dependence aggregation. The next step is to aggregate the interdependence between informative reference descriptors and each location of all input features $\{\mathbf{f}_n\}_{n=1}^{N}$. For convenience, we elaborate the process of dependence aggregation with one input feature $\mathbf{f}$, as such operation can be applied to other input features in parallel. A convolution layer is first applied to embed the input feature $\mathbf{f}$ into new feature map $\mathbf{V} \in \mathbb{R}^{C\times H\times W}$. We rewrite $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_{hw}]$, where $\mathbf{v}_i$ is a c-dimension column vector representing the local input feature. We aggregate interdependence by adaptively select the informative reference descriptors $\{\mathbf{g}_j\}_{j=1}^{c}$ to each location i of the input $\mathbf{f}$ conditioned on the need of local feature $\mathbf{v}_i$, obtaining attentive output denoted as $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_{hw}] \in \mathbb{R}^{C\times(HW)}$. For each position i, $\mathbf{z}_i$ can be computed as:

$$\mathbf{z}_i = \mathbf{G} \cdot softmax(\mathbf{v}_i)^\top = \sum_{j=1}^{c} \hat{v}_{i,j} \cdot \mathbf{g}_j, \tag{3}$$

$$\hat{v}_{i,j} = \frac{\exp(v_{i,j})}{\sum_j \exp(v_{i,j})}, \tag{4}$$

where $v_{i,j}$ indicates the j_{th} element of $\mathbf{v}_i$ and is normalized by softmax to generate $\hat{v}_{i,j}$. This soft attention operation is found to give better convergence.

Attentive pyramid feature. To regress a fine-grained depth map, we employ a pyramid-like structure to capture the multi-level interdependence at different scales to progressively build the cost volumes of higher resolutions. The LAM module applied at different levels generate multiple attentive features $\{\mathbf{Z}_n^l\}_{l=1}^{3}$ that contain different levels of interdependence, while high-level information contributes to reconstruct reflective or occluded regions, and low-level information helps to localize a pixel precisely.

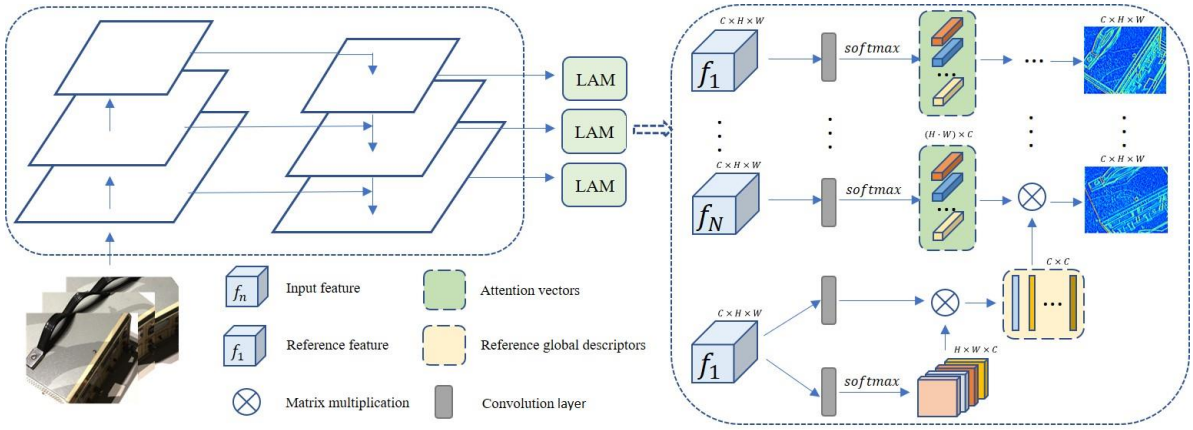

Figure 4. Architecture of the attentive pyramid feature extractor. We leverage the Feature Pyramid Network (FPN) to extract multi-scale features, and then feed them into our designed long range attention module (LAM). LAM first generates a set of reference global descriptors by second-order pooling. Then LAM selectively distributes them to each location of input via attention vectors to establish interdependence between reference and other input.

3.3. Depth prediction

Once we obtain the attentive features, we first warp them to construct the plane sweep volumes, which are then fused together to build the 3D cost volume similar to [14]. Compared with the previous U-Net, we replace the last two layers of the encoder architecture with the ASPP module to regularize the cost volume incorporating contextual information for depth inference. ASPP module enables the network to effectively enlarge the receptive fields to incorporate long-range context. Note that the architecture of ASPP in three stages is the same but without sharing weights, separately processing multi-scale cost volumes. The end layer of 3D convolutional network is a depth-wise softmax to predict per-pixel depth probability volume $\mathbf{P}$. For a given level l, assume that the depth hypotheses at pixel x are $\mathbf{d}^l(x)$ ranging from d_1^l to d_K^l, where K denotes the number of depth planes, and $\mathbf{P}_k^l(x)$ represents at pixel x how probable the depth is d_k^l. Consequently, we compute the estimated depth of each pixel x at the l_{th} level utilizing $expectation$ value along the depth direction:

$$\mathbf{D}^l(x) = \sum_{k=1}^{K} d_k^l(x) \cdot \mathbf{P}_k^l(x). \tag{5}$$

3.4. Combined loss function

Previous works train the network with the L1 depth loss, which measures the absolute difference between the predicted depth map and the ground truth depth map. However, it ignores that the probability volume is indirectly supervised as an intermediate layer, which leaves its distribution less constrained. To address this problem, we propose a new probability volume loss, coupled with the depth loss to train the network effectively.

Probability volume loss. Assuming at pixel x, the depth hypotheses are ranging from d_1 to d_K and the ground truth depth is $\hat{d}$. Probability volume $\mathbf{P}_k(x)$ represents at pixel x how probable the depth is d_k. Therefore, the predicted probability volume should have the highest probability at depth $\hat{d}$ and a gradually decreasing probability as the further away from the true depth. This property requires the probability distribution to peak at the true depth at each position in the ground truth probability volume, obeying a Gaussian-like distribution whose $expectation$ is $\hat{d}$. The ground truth probability volume at pixel x is defined as:

$$\hat{\mathbf{P}}_k(x) = \frac{\exp\left(\hat{c}_k(x)\right)}{\sum_{k=1}^{K} \exp\left(\hat{c}_k(x)\right)}, \tag{6}$$

$$\hat{c}_k(x) = -\frac{|d_k(x) - \hat{d}(x)|}{\sigma(x)}, \tag{7}$$

where $\hat{c}_k(x)$ indicates the normalized distance between depth hypothesis $d_k(x)$ and truth depth and σ is the variance that controls the sharpness of the peak around the true depth. Different pixels should have various sharpness. For example, those certainly matched pixels should have a sharp peak while those uncertain ones have a relatively flat peak.

To build such more reasonable labels for probability volume, we leverage a confidence score map $\mathbf{s} \in [0,1]^{H \times W}$ to adaptively predict σ for each pixel. Large score s means matching confidently and small score s represents matching ambiguously. Therefore, we can take the single peak of the probability distribution to measure the estimation confidence. Then, σ for generating ground truth distribution is computed as:

$$s(x) = max\{\mathbf{P}_k(x) \mid k \in [1, K]\}, \tag{8}$$

$$\sigma(x) = \alpha \cdot (1 - s(x)) + \beta, \tag{9}$$

where α is a scale factor that reflects the sensitivity of σ to the change of confidence s, and β defines the lower bound for σ and avoids numerical issue of dividing 0.

At pixel position x, we have obtained both estimated probability volume $\mathbf{P}_k(x)$ and the ground truth $\hat{\mathbf{P}}_k(x)$. The probability volume loss can be defined via cross entropy:

$$L_p = -\sum_{x \in \Omega} \sum_{k=1}^{K} \hat{\mathbf{P}}_k(x) \cdot log \mathbf{P}_k(x), \tag{10}$$

where Ω denotes the set of valid ground truth pixels.

Depth loss. Similar to existing MVSNet [34] framework, we also make use of the L1 norm measuring the absolute difference between the ground truth $\hat{\mathbf{D}}$ and the estimated depth $\mathbf{D}$ to define the depth loss as:

$$L_d = \sum_{x \in \Omega} ||\mathbf{D}(x) - \hat{\mathbf{D}}(x)||_1. \tag{11}$$

Combined loss. We apply the loss function to supervise all the three stages with a set of stage weight $\{\gamma^l\}_{l=1}^3$. In each stage, the combined loss is formulated as a weighted sum of depth loss L_d^l and probability volume loss L_p^l. The total loss is defined as:

$$L = \sum_{l=1}^{3} \gamma^l \cdot (L_d^l + \lambda \cdot L_p^l), \tag{12}$$

where λ is the loss weight to balance the contributions of depth loss and probability volume loss.

4. Experiments

In this section, we evaluate our proposed network on DTU dataset [1] and Tanks and Temples dataset [18]. First, we describe the datasets and evaluation metrics followed by implementation details. Then, we show the benchmarking results on the above datasets. At last, we perform ablation studies using DTU dataset.

4.1. Datasets and evaluation metrics

DTU [1] is a large-scale indoor MVS dataset consisting of 124 different scenes scanned in 7 different lighting conditions from 49 or 64 views. Each view consists of an image and the calibrated camera parameters. The size of images is 1600×1200, and the depth range of a scene is between $425mm$ and $935mm$. Point clouds with normal information are provided so that ground truth depth maps can be generated. We use the same training, validation and evaluation sets as defined in [34].

Tanks and Temples [18] is a large outdoor dataset captured in more complex environments. The dataset contains two sets of scenes, the intermediate one and the advanced one. We only use the intermediate set for evaluation which consists of eight scenes: Family, Francis, Horse, Lighthouse, M60, Panther, Playground, and Train. Their ground truth camera poses and ground truth point clouds are withheld by the evaluation website.

Methods	Acc.(mm)	Comp.(mm)	Overall(mm)
Camp [3]	0.835	0.554	0.695
Furu[12]	0.613	0.941	0.777
Tola [29]	0.342	1.190	0.766
Gipuma [13]	**0.283**	0.873	0.578
SurfaceNet [28]	0.450	1.040	0.745
MVSNet [34]	0.396	0.527	0.462
RMVSNet [35]	0.383	0.452	0.417
MVSCRF[32]	0.371	0.426	0.398
Point-MVSNet[7]	0.342	0.411	0.376
CVP [33]	0.296	0.406	0.351
UCS-Net [9]	0.330	0.372	0.351
Baseline [14]	0.346	0.351	0.348
LANet (Ours)	0.320	**0.349**	**0.335**

Table 1. Performance comparison on DTU dataset (lower means better).

Evaluation metrics. We take the commonly used evaluation metrics, accuracy and completeness in millimeter, to benchmark with previous methods. Accuracy is measured as the distance from estimated point clouds to the ground truth ones and completeness is defined as the distance from ground truth point clouds to the estimated ones. As for DTU dataset, the overall score is the average of accuracy and completeness. And for Tanks and Temples dataset, the F-score is used as the evaluation metric to measure the accuracy and completeness.

4.2. Implementation details

Training. We train our LANet on the DTU dataset. The backbone architecture employed in our network is the CasMVSNet [14]. During training we set input image resolution to 640×512 and the number of views $N = 3$. With the same coarse-to-fine manner as in CasMVSNet, for the first stage, the number of depth planes K is set to 48, which are uniformly sampled from $425mm$ to $921mm$. From the second to the third stage, depth planes K are set to 32 and 8, and the corresponding depth intervals are $5.3mm$ and $2.65mm$, respectively. Accordingly, the spacial resolutions of feature maps are $\{1/16,1/4,1\}$ of the input image size. The corresponding stage weights $\{\gamma^l\}_{l=1}^3$ are set to $\{0.5,1,2\}$ to balance the loss function following [14] that later stage prefers a larger stage weight. We implement our network using Pytorch [25]. The model is trained end-to-end by the Adam optimizer [17] with batch size of 16 for 16 epochs on 8 NVIDIA GTX 1080Ti GPUs. The initial learning rate is $1e-3$ and divided by 2 iteratively at the 10^{th},12^{th} and 14^{th} epoch.

Evaluation. As for DTU testing dataset, we follow [34] to fuse all predicted depth maps into point clouds using the

Figure 5. Qualitative results of reconstructed point clouds on benchmarks. Top row: results on DTU dataset [1]. Bottom row: results on Tanks and Temples dataset [18].

Methods	Mean	Family	Francis	Horse	Lighthouse	M60	Panther	Playground	Train
COLMAP [27, 26]	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04
MVSNet [34]	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
RMVSNet [35]	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
MVSCRF[32]	45.73	59.83	30.60	29.93	51.15	50.61	51.45	52.60	39.68
Point-MVSNet[7]	48.27	61.79	41.15	34.20	50.79	51.97	50.85	52.38	43.06
CVP [33]	54.03	76.5	47.74	36.34	**55.12**	**57.28**	**54.28**	**57.43**	47.54
UCS-Net [9]	53.14	70.93	51.75	42.66	53.43	54.33	50.67	54.37	47.02
Baseline [14]	54.57	**76.36**	52.79	47.02	54.02	54.15	50.24	51.45	50.49
LANet (Ours)	**55.70**	76.24	**54.32**	**49.85**	54.03	56.08	50.82	53.71	**50.57**

Table 2. Quantitative results of F-scores (higher means better) on Tanks and Temples dataset

same post-processing method. For fair comparisons, we use the same view selection $N = 5$ and imgae size 1600×1184. To validate the generation of our method, we also test it on the intermediate set of Tanks and Temples dataset using the model trained on DTU dataset without fine-tuning.

4.3. Results on DTU dataset

We compare our proposed LANet with the state-of-the-art including both traditional methods and leaning-based methods. As shown in Table 1, our LANet significantly outperforms all previous models in term of completeness and overall scores while traditional method Gipuma [13] achieves the best accuracy. Specifically, LANet achieves $0.320mm$ and $0.349mm$ in terms of accuracy and completeness, respectively. LANet outperforms the previous best performing method [14] by 3.7% in overall scores, which demonstrates the great benefit of our proposed method. Fig. 2 and Fig. 5 show some qualitative point cloud results. As shown in Fig. 2, our method is able to obtain more complete in texture-less area than RMVSNet [35] and CasMVSNet [14]. For instance, in the regions highlighted by the red box, our results have smoother surfaces and clearer boundaries. Note that, we get the point cloud results of above mentioned methods by running their provided pretrained model and code except RMVSNet which provides point cloud reconstruction on its website.

4.4. Results on Tanks and Temples dataset

To evaluate the generalization of LANet, we use the model trained on DTU dataset without any fine-tuning to reconstruct point clouds in Tanks and Temples intermediate dataset. Similar to previous works, we leverage the same view selection, image size, camera parameters and initial depth range as in MVSNet [34] with $N = 5$, $H = 1056$, $W = 1920$. The reconstructed point clouds are submitted to the evaluation website [18] to get the quantitative results of F-scores. As shown in Table 2, our LANet achieves the best result, which is significantly better than the baseline model [14] by 1.13% in terms of mean F-scores. In particular, for scene Horse and Train, our method obtains the new state-of-the-art performance. Note that, since [14] has not released the source code for depth map fusion on Tanks and Temples dataset, we reproduce this baseline model following the commonly used fusion process provided by MVSNet. Some reconstructed 3D point clouds by our method are shown in Fig. 5 to demonstrate the quality of the recon-

Methods	Acc.(mm)	Comp.(mm)	Overall(mm)
Baseline	0.346	0.351	0.348
w/ LAM	0.323	0.358	0.340
w/ ASPP	0.341	0.350	0.346
w/ PVL	0.348	0.337	0.342
Full ver.	**0.320**	**0.349**	**0.335**

Table 3. Quantitative evaluation of each component.Baseline refers to the CasMVSNet pipeline. LAM refers to the long-range attention module. ASPP refers to replace 3D U-Net with 3D ASPP. PVL refers probability volume loss. Full ver. means including all the components.

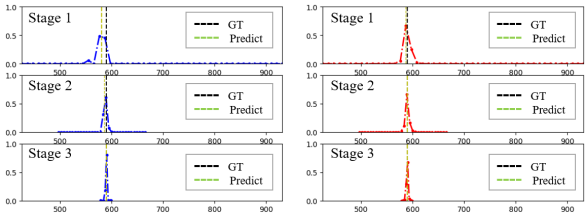

Figure 6. Comparison of predicted probability volume distribution at three stages. The left column is the output of CasMVSNet [14] and the right column is the output of our LANet. Note that the black dash line corresponds to the ground truth depth and the yellow dash line corresponds to the predicted depth.

struction.

4.5. Ablation study

To gain insight into the proposed LANet, we perform an ablation study to demonstrate the contribution of each component. We follow the previous works [14, 22, 7], using DTU dataset as the benchmark for the ablation study.

Benefit of each component. To illustrate the improvement brought by each component, we first implement a baseline model used in CasMVSNet and then compare it with our LANet on each component. As shown in Table 3, with the LAM only, the accuracy and overall score are reduced significantly by $0.023mm$ and $0.008mm$, respectively.The results demonstrate the effectiveness of the proposed attention module on capturing long-range interdependence of multi-view images. Further, the ASPP module reduces the overall score by $0.002mm$. This indicates that our 3D ASPP enables the network to incorporate long-range context information for effective cost regularization.

The results of PVL in Table 3 show that with probability volume loss, the performance has been largely improved by $0.014mm$ in terms of completeness. Visual comparisons are shown in Fig. 6, for those falsely matched pixels in CasMVSNet, their probability distributions are scattered and cannot be centralized to the ground truth especially in the first stage. By leveraging the intermediate supervision, the $expectation$ of our predicted distribution is close to the truth depth.

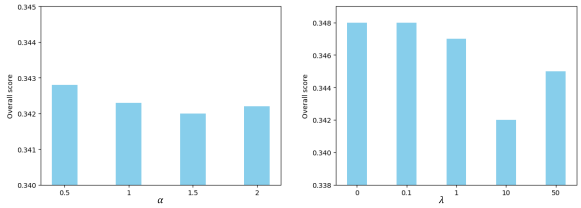

Figure 7. Influence of sensitivity α and loss weight λ for the final prediction performance.

Probability volume distribution. The shape of probability volume distribution is mainly adjusted by variance σ, where $\sigma \in [\beta, \alpha+\beta]$ is bounded by hyperparameters α and β shown in Eq.9. We first study the case when the variance σ is fixed for all pixels and we find $\sigma = 0.5$ performs the best by grid searching.This indicates that most pixels favor small variances and sharp distributions. Therefore, we set the lower bound β to 0.1 to avoid numerical issue of dividing 0 and then analyze how sensitivity α affects the adaptive variance study. As shown in the left picture of Fig. 7, when $\alpha = 1.5$ the performance is best and the result is relatively stable by varying α from 0.5 to 2.0.

Loss weight. We conduct experiments with various combinations of loss weights on DTU dataset. As illustrated in the right picture of Fig. 7, the combinatorial loss with both depth loss and probability volume loss shows better performance while loss weight $\lambda = 10$ stands out, with $0.006mm$ reduced in terms of overall scores.

5. Conclusion

In this paper, we introduce a Long-range Attention Network for MVS to capture the interdependence among pixels, propagating more guiding information to measure the similarity between the corresponding image pairs. Furthermore, we introduce a new loss to supervise the distribution of probability volumes to improve its robustness for depth inference. Extensive experiments on DTU and Tanks & Temples benchmarks show our LANet outperforms previous state-of-the-art methods. Thorough ablation studies demonstrate the benefits of the introduced long-range attention module and the probability volume loss for MVS.

Acknowledgments

This work was supported by the National Security Major Basic Research Program of China under Grant 15001303, the National Key Scientific Instrument and Equipment Development Project under Grant 61827901, and the Guangxi Municipal Science and Technology Project under Grant 31062501. This study was also supported by Grant NO.2019JZZY011101 from the Key Research and Development Program of Shandong Province to Dianmin Sun.

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(2):153–168, 2016.

[2] Konstantinos Batsos, Changjiang Cai, and Philippos Mordohai. Cbmv: A coalesced bidirectional matching volume for disparity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2060–2069, 2018.

[3] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, pages 766–779. Springer, 2008.

[4] Joao Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *European Conference on Computer Vision*, pages 430–443. Springer, 2012.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[7] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1538–1547, 2019.

[8] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Aˆ 2-nets: Double attention networks. In *Advances in Neural Information Processing Systems*, pages 352–361, 2018.

[9] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020.

[10] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[11] Simon Fuhrmann, Fabian Langguth, and Michael Goesele. Mve-a multi-view reconstruction environment. In *GCH*, pages 11–18. Citeseer, 2014.

[12] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

[13] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015.

[14] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020.

[15] Yutao Hu, Yandan Yang, Jun Zhang, Xianbin Cao, and Xiantong Zhen. Attentional kernel encoding networks for fine-grained visual categorization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[16] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in neural information processing systems*, pages 365–376, 2017.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] 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.

[19] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Shape reconstruction using volume sweeping and learned photoconsistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 781–796, 2018.

[20] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[21] Peihua Li, Jiangtao Xie, Qilong Wang, and Wangmeng Zuo. Is second-order information helpful for large-scale visual recognition? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2070–2078, 2017.

[22] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10452–10461, 2019.

[23] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5695–5703, 2016.

[24] Despoina Paschalidou, Osman Ulusoy, Carolin Schmitt, Luc Van Gool, and Andreas Geiger. Raynet: Learning volumetric 3d reconstruction with ray potentials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2018.

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[26] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[27] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016.

[28] Ayan Sinha, Asim Unmesh, Qixing Huang, and Karthik Ramani. Surfnet: Generating 3d shape surfaces using deep residual networks. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 6040–6049, 2017.

[29] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23(5):903–920, 2012.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[31] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[32] Youze Xue, Jiansheng Chen, Weitao Wan, Yiqing Huang, Cheng Yu, Tianpeng Li, and Jiayu Bao. Mvscrf: Learning multi-view stereo with conditional random fields. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4312–4321, 2019.

[33] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020.

[34] 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)*, pages 767–783, 2018.

[35] 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 Conference on Computer Vision and Pattern Recognition*, pages 5525–5534, 2019.

[36] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[37] Youmin Zhang, Yimin Chen, Xiao Bai, Jun Zhou, Kun Yu, Zhiwei Li, and Kuiyuan Yang. Adaptive unimodal cost volume filtering for deep stereo matching. *arXiv preprint arXiv:1909.03751*, 2019.