Learnable Cost Metric-Based Multi-View Stereo for Point Cloud Reconstruction

Guidong Yang, *Graduate Student Member, IEEE*, Xunkuai Zhou, Chuanxiang Gao, Xi Chen, and Ben M. Chen, *Fellow, IEEE*

***Abstract*—3-D reconstruction is essential to defect localization. This article proposes LCM-MVSNet, a novel multi-view stereo (MVS) network with learnable cost metric (LCM) for more accurate and complete dense point cloud reconstruction. To adapt to the scene variation and improve the reconstruction quality in non-Lambertian low-textured scenes, we propose LCM to adaptively aggregate multi-view matching similarity into the 3-D cost volume by leveraging sparse point hints. The proposed LCM benefits the MVS approaches in four folds, including depth estimation enhancement, reconstruction quality improvement, memory footprint reduction, and computational burden alleviation, allowing the depth inference for high-resolution images to achieve more accurate and complete reconstruction. In addition, we improve the depth estimation by enhancing the shallow feature propagation via a bottom–up pathway and strengthen the end-to-end supervision by adapting the focal loss to reduce ambiguity caused by sample imbalance. Extensive experiments on three benchmark datasets show that our method achieves state-of-the-art performance on the *DTU* and *BlendedMVS* dataset, and exhibits strong generalization ability with a competitive performance on the *Tanks and Temples* benchmark. Furthermore, we deploy our LCM-MVSNet into our UAV-based infrastructure defect inspection framework for infrastructure reconstruction and defect localization, demonstrating the effectiveness and efficiency of our method. More experiment results can be found in the Appendix at https://github.com/CUHK-USR-Group/TIE_Appendices/blob/main/TIE_Appendix.pdf.**

***Index Terms*—Defect inspection, depth estimation, diagnosis and monitoring, intelligent system, multi-view stereo (MVS), reconstruction, unmanned aerial vehicle (UAV).**

Manuscript received 15 April 2023; revised 13 August 2023 and 1 October 2023; accepted 15 November 2023. Date of publication 14 December 2023; date of current version 5 June 2024. This work was supported in part by the InnoHK of the Government of the Hong Kong Special Administrative Region via the Hong Kong Centre for Logistics Robotics, and in part by the Research Grants Council of Hong Kong SAR under Grant 14217922. *(Corresponding author: Guidong Yang.)*

The authors are with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: gdyang@mae.cuhk.edu.hk; xunkuaizhou@cuhk.edu.hk; cxgao@mae.cuhk.edu.hk; xichen@mae.cuhk.edu.hk; bmchen@cuhk.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIE.2023.3337697.

Digital Object Identifier 10.1109/TIE.2023.3337697

I. Introduction and Literature Review

MULTI-VIEW stereo (MVS) aims to recover the dense 3-D representation of the scene leveraging stereo correspondences as the main cue given calibrated 2-D images from multiple views (more than two views), essentially equivalent to solving the pixel correspondences across multi-view images. Recently, learning-based MVS approaches [1], [2], [3], [4], [5], [6], [7], [8], [9] have significantly outperformed the traditional counterparts in MVS benchmarks [10], [11], [12], [13]. Deep MVS approaches decouple the MVS into a two-stage process: *learning-based depth map estimation* and *depth map filtering and fusion*. Compared to the handcrafted photometric measures in traditional approaches, deep MVS approaches encode scene cues, such as reflective priors and illumination changes into the network by adopting powerful feature extraction and cost volume representation to achieve superior reconstruction accuracy and completeness. Despite the superiority of the learning-based MVS approaches, the following improvements can be made to further boost the overall reconstruction quality.

Most learning-based methods [1], [2], [3], [4], [5], [6], [7], [8], [9] use feature pyramid network (FPN) to extract multiscale features for constructing cost volume pyramid. We observe that these methods suffer from oversmoothing depth estimation (see Fig. 5 of the Appendix) around the object boundaries due to the lack of shallow feature information containing low-level features, such as local textures and edges. To tackle this issue, we introduce a bottom–up pathway with negligible parameter increases to shorten the propagation of shallow information, shown to be conducive to both depth estimation and reconstruction.

Based on multi-view deep features, learning-based methods construct the cost volume [14] to encode scene context and geometries into the network, where the cost volume is a 3-D volume (depth$\times$ height $\times$ width) measuring the multi-view feature matching cost between the reference and source-view feature maps along the depth. The cost volume is then regularized to produce the depth map estimation. Effective cost volume aggregation is crucial to ensure multi-view photo consistency. Recent learning-based MVS approaches adopt two types of schemes for cost volume aggregation: heuristic and learning-based. The heuristic cost volume aggregation [1], [15] assigns equal significance to each of the multi-view feature volumes to aggregate the cost volume, ignoring the scene variation among different views. The learning-based cost volume aggregation generally applies

an additional reweight network to learn the pixelwise weight [4], patchwise weight [16], channelwise weight [17], or voxelwise weight [5], [8], [18] for cost volume aggregation. However, the additional reweight network imposes a computational burden and ignores the intrinsic correspondences between multi-view images. To address the aforementioned limitations, we propose the *learnable cost metric* (LCM) scheme to achieve a tradeoff between heuristic and learning-based cost volume aggregation. Heuristic aggregation methods ignore the scene variation among different views. Nevertheless, we observe that images from different views have pixel differences caused by illumination changes, occlusions, and image content variations. Moreover, the source image near the reference view without occlusion can offer more accurate photometric and geometric information than a far one with partial occlusion. Based on this observation, LCM computes the per-view significance to account for per-view scene variation. To alleviate the memory consumption and computational burden imposed by learning-based aggregation methods, LCM scheme introduces the sparse point hints of the scene from SfM into the aggregation process to directly compute the source-view significance and learn the reference-view significance from the training data. The resulting LCM scheme adapts to multi-view scene variation and concurrently alleviates the computational burden. Abiding by the LCM scheme, we present two LCM modules including *coarse-to-fine LCM* and *efficient LCM*, both effectively improving the accuracy and completeness of the depth estimation and reconstruction, while the former one is more accurate and the later one is more efficient. Meanwhile, the proposed modules are adaptive to an arbitrary number of input views, independent of the order of input views, scalable to large-scale scenarios, and complementary to cost volume-based MVS approaches.

Most recent learning-based MVS methods adopt two types of loss functions for end-to-end training: L_1 loss [1], [3], [7] and cross entropy loss [2], [19], [20]. We find that the regression-based L_1 loss represents the mean absolute error between the ground truth and regressed depth values, leading to imbalance between the single target depth and multiple possible sets of weight combination. And the classification-based cross entropy loss supervises the cross entropy between the ground truth and classified probabilities, resulting in discretized depth estimations. To alleviate the imbalance and achieve continuous depth estimation, we adapt the *focal* loss [8], [9], [21], [22] in the object detection field to the MVS task to supervise the cross entropy between the estimated depth bias and ground-truth depth bias in a continuous manner, shown to achieve more accurate and complete depth estimation and dense point cloud reconstruction.

To this end, we present a novel MVS network with LCM for depth estimation, termed as LCM-MVSNet. Our network follows the coarse-to-fine framework [3] to further boost the depth estimation and reconstruction performance, concurrently reduce the memory footprint and speed up the inference. The network takes as input the multi-view images and outputs per-view depth map pyramid, from which the depth map at the finest level is taken as the final output. Multi-View depth maps filtering and fusion are then conducted to obtain the dense point cloud reconstruction. Extensive experiments show that our method achieves state-of-the-art performance on the *DTU* [10] and *BlendedMVS* [13] datasets, and exhibits strong generalization ability on the *Tanks and Temples* [11] benchmark. Systematic ablation experiments further verify the effectiveness of each component of our method.

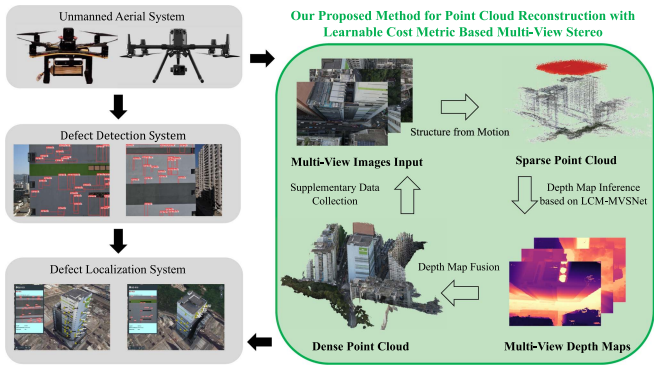


Fig. 1. Our UAV-based infrastructure defect inspection framework for defect localization.

To show the practicality and robustness of our method, we deploy our method into our proposed UAV-based infrastructure defect inspection framework for infrastructure reconstruction and defect localization, with crack as our research target. As shown in Fig. 1, our inspection framework comprises *unmanned aerial system* (UAS), *defect detection system* [23], proposed MVS method (**green** block), and *defect localization system*. The UAS with advanced motion planning and control algorithms is developed to autonomously collect close-range high-resolution visual images of the target infrastructure. Then, the images are processed via the *defect detection system* for defect identification and the proposed MVS method for 3-D reconstruction, respectively. Finally, the detected defects are localized and registered into the reconstructed 3-D model of the target infrastructure through geographic information. Our UAV-based infrastructure defect inspection framework applies to multiscale scenarios due to the flexible scalability of our proposed MVS method.

The rest of this article is organized as follows. Our MVS methodology is illustrated in Section II, followed by Section III detailing the experiments on the standard MVS benchmarks. Section IV presents the real-world application for UAV-based infrastructure defect inspection. Section V discusses the limitations of the proposed method. Finally, Section VI concludes this article.

II. Methodology of MVS

In this section, we illustrate our two-stage MVS method for automated point cloud reconstruction: LCM-MVSNet for depth inference, depth map filtering and fusion strategy for the point cloud generation. Fig. 2 depicts the outline of the network, taking as input multi-view images $\{\mathbf{I}_i \in \mathbb{R}^{C\times H\times W}\}_{i=0}^{N}$ of the scene from $N+1$ views with known camera parameters, where $\mathbf{I}_0$ represents the reference image, $\{\mathbf{I}_i\}_{i=1}^{N}$ denotes N neighboring source images, and C, H, W is the channel number and spatial dimension of the input image, respectively. The network infers the depth map $\mathbf{D}_0$ for the reference image $\mathbf{I}_0$, where each image $\mathbf{I}_i$ from $\{\mathbf{I}_i\}_{i=0}^{N}$ is iteratively treated as the reference image. The estimated depth maps $\{\mathbf{D}_i\}_{i=0}^{N}$ are then filtered and fused

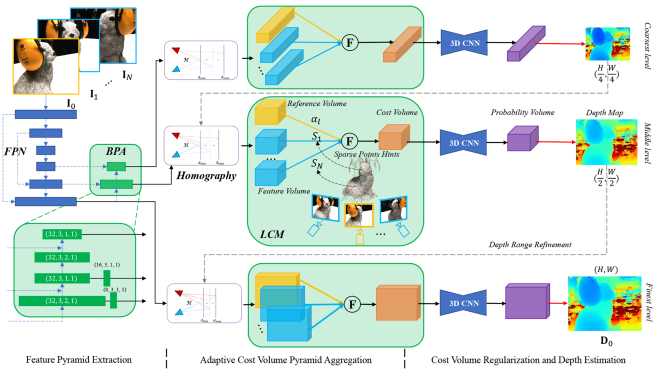


Fig. 2. Overview of our proposed LCM-MVSNet (see Appendix for best view).

to generate the point cloud reconstruction. In the following, we successively detail each component of our method: feature pyramid extraction, adaptive cost volume pyramid aggregation, cost volume regularization and depth estimation, loss function for optimization, and depth map filtering and fusion strategy.

A. Feature Pyramid Extraction

Most recent learning-based methods [7], [8], [9] adopts the coarse-to-fine depth estimation strategy and utilizes FPN to extract multiscale image features for constructing cost volumes at different resolutions. We observe that these methods suffer from oversmoothing depth estimation (see Fig. 5 of the Appendix) around the object boundaries due to the lack of shallow feature information containing low-level features, such as local textures and edges, we hence enhance the shallow feature information flow by introducing a bottom–up pathway to augment the propagation of low-level features and enlarge the receptive field to incorporate global context information for more accurate and robust feature matching under low-textured regions. Our feature pyramid extraction network takes as input multi-view images $\{\mathbf{I}_i\}_{i=0}^{N}$ and output $(L+1)$-level feature pyramids $\{\mathbf{f}_{l,i} \in \mathbb{R}^{F_l \times H/2^l \times W/2^l}\}_{l=0}^{L}$ for each image $\mathbf{I}_i$, where l represents the level ordinal, $l = L$ is the coarsest level, $l = 0$ is the finest level, F_l is the channel number of the feature map at level l, $H/2^l$ and $W/2^l$ is the height and width of the lth level feature map downsampled to $1/2^l$ of the original input image resolution, respectively.

Specifically, the network comprises 20 convolutional layers, including 16 layers for FPN [3], and four layers for bottom–up path augmentation (BPA), please refer to Fig. 2 for detailed layer settings of BPA where (A, B, C, D) denotes channel number, kernel size, stride, and padding size, respectively. We replace batch normalization with in-place activated batch normalization (ABN) to reduce memory footprint in a computationally efficient way [24]. L is set to 3 to construct a 3-level (3-stage) feature extraction network. The spatial resolution of feature pyramid at level $l = 0, 1, 2$ is $H \times W$, $H/2 \times W/2$, and $H/4 \times W/4$, respectively. F_l is set to $8, 16, 32$ for $l = 0, 1, 2$, respectively, to build up the F_l-level pixel descriptors for encoding the original neighboring information, preventing the subsequent dense matching from losing useful context information. Systematic ablation experiments in Section III-C show that the BPA can improve the depth estimation and reconstruction quality.

B. Adaptive Cost Volume Pyramid Aggregation

The next step is to encode the extracted image features $\{\mathbf{f}_{l,i} \in \mathbb{R}^{F_l \times H/2^l \times W/2^l}\}_{i=0}^{N}$ of $(N+1)$-view images $\{\mathbf{I}_i\}_{i=0}^{N}$ and camera parameters into the network for multi-view feature volumes construction and cost volume aggregation.

For level l, we uniformly sample $(M_l + 1)$-layer depth hypotheses in 3-D space from the depth range $[d_{\min,l}, d_{\max,l}]$ for the reference camera frustum

$$d_{m,l} = d_{\min,l} + m\frac{d_{\max,l} - d_{\min,l}}{M_l} \tag{1}$$

where $d_{\min,l}, d_{\max,l}$ represents the minimum and maximum depth at level l, respectively, $m \in \{0, 1, \ldots, M_l\}$ stands for the sample index of depth hypothesis where its normal vector $\mathbf{n}_0$ is the principal axis of the reference camera, and $M_l + 1$ is the total sample number of depth hypotheses at level l. Notably, the depth range $[d_{\min,L}, d_{\max,L}]$ of the coarsest level $l = L$ is predefined and the depth range $[d_{\min,l}, d_{\max,l}]$ for the finer level is dynamically determined by centering the depth range at the depth estimation (detailed in Section II-C) of the previous level, concurrently reducing depth interval and sample number [3]. Please refer to Section III-B of the Appendix for the coarse-to-fine depth estimation strategy.

With sampled depth hypotheses, source-view feature volumes are constructed in the 3-D space through differentiable homography by warping the extracted 2-D source-view image features into the reference camera frustum. For level l, the homography matrix $\mathbf{H}_i(d_{m,l})$ between ith source-view feature map and reference feature map at depth $d_{m,l}$ is defined as follows:

$$\mathbf{H}_i(d_{m,l}) = \mathbf{K}_i^l \mathbf{R}_i \left(\mathbf{I} - \frac{(\mathbf{C}_0 - \mathbf{C}_i)\mathbf{n}_0^T}{d_{m,l}}\right) \mathbf{R}_0^T (\mathbf{K}_0^l)^{-1} \tag{2}$$

where $\mathbf{I} \in \mathbb{R}^{3\times3}$ is the identity matrix, $\{\mathbf{K}_i^l, \mathbf{R}_i\} \in \mathbb{R}^{3\times3}$ refers to the scaled camera intrinsic at level l and rotation matrix of the ith source view respectively, and $\mathbf{C}_i \in \mathbb{R}^{3\times1}$ is inhomogeneous coordinates of the camera center of the ith source view. $\mathbf{K}_0^l, \mathbf{R}_0, \mathbf{C}_0$ refer to counterparts of the reference view. $\mathbf{n}_0$ denotes the principal axis of the reference camera and T stands for the matrix transpose. The warping process is then achieved by differentiable bilinear interpolation to sample the source-view image features $\{\mathbf{f}_{l,i} \in \mathbb{R}^{F_l \times H/2^l \times W/2^l}\}_{i=1}^{N}$ into the reference view to generate source-view feature volumes $\{\mathbf{V}_{l,i} \in \mathbb{R}^{F_l \times M_l \times H/2^l \times W/2^l}\}_{i=1}^{N}$. The reference volume $\mathbf{V}_{l,0} \in \mathbb{R}^{F_l \times M_l \times H/2^l \times W/2^l}$ is acquired by repeating the reference-view image features $\mathbf{f}_{l,0} \in \mathbb{R}^{F_l \times H/2^l \times W/2^l}$ M_l times. Please refer to Section III-C of the Appendix for the graphic demonstration of the cost volume construction.

To adapt to the arbitrary number of input views, the next step is to aggregate $(N+1)$-view volumes $\{\mathbf{V}_{l,i} \in \mathbb{R}^{F_l \times M_l \times H/2^l \times W/2^l}\}_{i=0}^{N}$ into a single cost volume $\mathbf{C} \in \mathbb{R}^{F_{l,c} \times M_l \times H/2^l \times W/2^l}$ for multi-view feature matching similarity measurement. This process can be defined as a

mapping function $\mathcal{M}: \underbrace{\mathbb{R}^{V_l} \times \cdots \times \mathbb{R}^{V_l}}_{\mathrm{N}} \rightarrow \mathbb{R}^{V_{l,c}}$, where $V_l = F_l \times M_l \times H/2^l \times W/2^l$ and $V_{l,c} = F_{l,c} \times M_l \times H/2^l \times W/2^l$. The heuristic cost volume aggregation scheme assigns equal significance to the reference view and each source view when aggregating cost volume, with the assumption that volumes of all views contribute equally to the 3-D cost volume. Nonetheless, we observe that images from different views have pixel differences caused by illumination changes, occlusions, and image content variations. Thus, feature volumes from different views should contribute differently to the cost volume aggregation. As the depth map of the reference image needs to be inferred, its image features should be critical to the cost volume aggregation process. Furthermore, the source image near the reference view without occlusion can offer more accurate photometric and geometric information than a far one with partial occlusion. Based on this important observation, we propose the *learnable cost metric* (LCM) scheme for cost volume aggregation: the reference-view significance is learned from the training data and the corresponding normalized matching score is adopted as the source-view significance to make the network adaptive to the input scene variation, where the matching score measuring the feature similarity between the source image and reference image is computed by utilizing their common sparse points obtained through the structure from motion (SfM). In the following, we introduce two modules abiding by the LCM scheme: *coarse-to-fine LCM* and *efficient LCM*.

Coarse-to-fine LCM: The *coarse-to-fine LCM* module at network level l is defined as follows:

$$\begin{aligned}\mathbf{C}_l &= \mathcal{M}(\mathbf{V}_{l,0}, \ldots, \mathbf{V}_{l,N}) \\ &= \alpha_l(\mathbf{V}_{l,0} - \overline{\mathbf{V}_l})^2 + \sum_{i=1}^{N} \frac{S_i}{\sum_{i=1}^{N} S_i}(\mathbf{V}_{l,i} - \overline{\mathbf{V}_l})^2 \end{aligned} \tag{3}$$

where $\mathbf{C}_l \in \mathbb{R}^{F_l \times M_l \times H/2^l \times W/2^l}$ represents the cost volume, N represents the number of input views, $\alpha_l, l \in \{0, \ldots, L\}$ is the level-dependent learnable significance of the reference view. $\overline{\mathbf{V}_l}$ is the mean of $(N+1)$-view volumes $\{\mathbf{V}_{l,i}\}_{i=0}^{N}$. $\{S_i\}_{i=1}^{N}$ represents the matching score between ith source image $\{\mathbf{I}_i\}_{i=1}^{N}$ and the reference image $\mathbf{I}_0$, and $\{S_i\}_{i=1}^{N}$ is computed by leveraging common sparse points from SfM process, the computational procedure is detailed in the Appendix.

Specifically, *coarse-to-fine LCM* first compute the mean volume $\overline{\mathbf{V}_l}$ and measure the per-view feature differences $(\mathbf{V}_{l,i} - \overline{\mathbf{V}_l})^2$. Then, *coarse-to-fine LCM* abides by the coarse-to-fine framework by setting different learnable a_l at different level $l \in \{0, 1, \ldots, L\}$ of the network, expecting the network to learn the significance of the reference-view feature differences from the training data. The level-dependent a_l enables the cost volume aggregation to adapt to network levels for more accurate multi-view matching similarity measurement. We set the normalized matching score $\frac{S_i}{\sum_{i=1}^{N} S_i}$ as the significance of each source-view feature differences to make the network adaptive to the input scene variation. This adaptability is attributed to the intrinsic nature of the matching score $\{S_i\}_{i=1}^{N}$, as its computational process relies on the common sparse points hints between the reference image and source images. The common sparse points are obtained from the SfM process, taking the raw image representations of the scene into consideration where the illumination changes, occlusions and image content variations are better revealed. Consequently, the matching score $\{S_i\}_{i=1}^{N}$ varies and adapts to different input scenes, enhancing the generalization ability of the network by improving the depth estimation and overall reconstruction quality.

Efficient LCM: The *efficient LCM* at network level l is defined as follows:

$$\begin{aligned}\mathbf{C}_l &= \mathcal{M}(\mathbf{V}_{l,0}, \ldots, \mathbf{V}_{l,N}) \\ &= \mathcal{M}(\mathbf{B}_{l,0}, \ldots, \mathbf{B}_{l,N}) \\ &= \mathrm{AvgPool}\left(\alpha_l \mathbf{B}_{l,0} \odot \sum_{i=1}^{N} \frac{S_i}{\sum_{i=1}^{N} S_i} \mathbf{B}_{l,i}\right) \end{aligned} \tag{4}$$

where $\mathbf{B}_{l,i} \in \mathbb{R}^{K \times (F_l/K) \times M_l \times H/2^l \times W/2^l}$ stands for the batched volumes after evenly separating the original volumes $\mathbf{V}_{l,i}$ into K batches along the channel dimension. Similar to the *coarse-to-fine LCM*, we set different learnable α_l at different level $l \in \{0, 1, ..., L\}$ of the network to learn the reference-view significance and set the normalized matching score $\frac{S_i}{\sum_{i=1}^{N} S_i}$ as the source-view significance to adapt to the scene variation. Then, we adopt Hadamard product $\odot$ to merge multi-view weighted volumes and apply average pooling along the channel dimension to compute the multi-view feature matching similarity for obtaining the cost volume $\mathbf{C}_l \in \mathbb{R}^{K \times M_l \times H/2^l \times W/2^l}$. Compared to the square sum operation, the Hardmard product and average pooling can significantly reduce the memory footprint and speed up the inference time. By setting K as a small positive integer (such as 2, 4, 8), the channel number of the cost volume can be compressed, further shrinking the memory consumption and alleviate the computational burden both in the cost volume aggregation process and subsequent regularization based on 3-D CNN.

Systematic ablation experiments in Section III-C demonstrates the effectiveness and efficiency of the proposed *coarse-to-fine LCM* and *efficient LCM*, they both outperform the heuristic and learning-based cost volume aggregation schemes by achieving more accurate and complete depth estimation and point cloud reconstruction.

C. Cost Volume Regularization and Depth Estimation

Following recent learning-based MVS methods [3], [7], [9], a four-scale 3-D CNN is adopted to regularize the aggregated cost volume pyramid $\{\mathbf{C}_l\}_{l=0}^{L}$ and output probability volume pyramid $\{\mathbf{P}_{l,\mathrm{est}}\}_{l=0}^{L}$ through *sigmoid* activation. Here, we replace the *softmax* activation with *sigmoid* for greater numerical stability when combined with the focal loss [21], [22]. For level l, the per-pixel depth estimation is achieved via the *winner-take-all* operation to get the discrete depth estimation. To get continuous depth estimation, we further refine the discrete depth estimation with the estimated bias between the target depth and discretized depth

$$\mathbf{D}_{l,\mathrm{est}} = \underbrace{\underset{d_{m,l} \in [d_{\min,l}, d_{\max,l}]}{\arg\max} \mathbf{P}_{l,\mathrm{est}}(d_{m,l})}_{\text{discrete depth}}$$

$$+\underbrace{\underbrace{\frac{(d_{\max,l}-d_{\min,l})}{M_l}}_{\text{depth interval}}\underbrace{\max \mathbf{P}_{l,\text{est}}(d_{m,l})}_{\text{normalized bias}}}_{\text{estimated bias}} \tag{5}$$

where $\mathbf{P}_{l,\text{est}}(d_{m,l})$ is the probability map at depth hypothesis $d_{m,l}$, $\mathbf{D}_{l,\text{est}}$ is the depth estimation at level l. Note that the depth estimation $\mathbf{D}_0$ at the finest level is taken as the output.

D. Loss Function

Most recent learning-based MVS approaches [3], [7], [25] use *Soft-argmin* operation to learn a weight combination of discrete depth hypotheses and get the depth estimation. L_1 loss is adopted to minimize the mean absolute distance between regressed $\mathbf{D}_{\text{est}}$ and ground-truth depth $\mathbf{D}_{\text{gt}}$. However, the imbalance between the single target depth and multiple possible sets of weight combination imposes ambiguity to the learning process [8], [9]. To alleviate this ambiguity, we adapt focal loss [8], [9], [21], [22] into the MVS training to directly supervise on the probability volume by extending the binary cross entropy into its complete form and generalizing the scaling factor into the absolute difference between the estimated $\mathbf{P}_{\text{est}}$ and ground-truth probability volume $\mathbf{P}_{\text{gt}}$, defined as the normalized bias between the ground-truth depth and the discrete depth hypotheses. For network level l, the adapted focal loss is as follows:

$$\mathcal{L}_l = \sum_{\mathbf{x}\in\{\mathbf{x}_{\text{valid}}\}} -\beta_l |\mathbf{P}_{l,\text{gt}}(\mathbf{x}) - \mathbf{P}_{l,\text{est}}(\mathbf{x})|^{\gamma_l} \cdot ((1-\mathbf{P}_{l,\text{gt}}(\mathbf{x}))\log(1-\mathbf{P}_{l,\text{est}}(\mathbf{x})) + \mathbf{P}_{l,\text{gt}}(\mathbf{x})\log(\mathbf{P}_{l,\text{est}}(\mathbf{x}))) \tag{6}$$

where $\mathbf{P}_{l,\text{est}}(\mathbf{x})$ is the estimated probability volume at pixel $\mathbf{x}$ from set $\{\mathbf{x}_{\text{valid}}\}$ denoting the set of pixels with valid ground truth, β_l and γ_l is the tunable balancing and focusing parameter at level l, respectively. The total loss function is defined as the weighted sum of the per-level loss $\mathcal{L}_l$

$$\mathcal{L} = \sum_{l=0}^{L} \lambda_l \mathcal{L}_l \tag{7}$$

where λ_l denotes the loss weight at level l. Systematic ablation experiments in Section III-C show that the adapted focal loss effectively boosts the depth estimation and reconstruction quality.

E. Depth Maps Filtering and Fusion

Depth maps filtering and fusion are conducted to fuse the estimated multi-view depth maps $\{\mathbf{D}_i\}_{i=0}^{N}$ into the final 3-D point cloud. For depth map filtering, we impose photometric and geometric constraints by setting the probability threshold τ to discard depth outliers and the number of consistent views N_c to reduce the depth inconsistency, respectively, where the photometric constraint estimates the multi-view matching quality and geometric constraint represents the multi-view depth consistency. After filtering, we fuse the estimated depth maps into the final 3-D point cloud as previous works [7], [8], [9].

III. Experiments on the Benchmark Datasets

In this section, we first demonstrate the effectiveness and superiority of our LCM-MVSNet on multiple MVS benchmarks including: *DTU* [10], [26], *BlendedMVS* [13], and *Tanks and Temples* [11]. We then conduct extensive ablation experiments to verify the effectiveness and efficiency of the method components. Finally, we show the scalability of our method in terms of domain and scale adaptability.

A. Datasets and Evaluation Metrics

Datasets: We train our LCM-MVSNet on the *DTU training set* and then evaluate it on the *DTU evaluation set* for quantitative benchmarking of the reconstruction performance. To benchmark the reconstruction performance on the *Tanks and Temples*, we further fine-tune the trained model on the *BlendedMVS training set* with more complex scene variations and diverse camera trajectories to improve the generalization ability. As our network is dedicated to depth map estimation, we also benchmark the depth map estimation quality on the *BlendedMVS validation set*.

Evaluation metrics: 1) Reconstruction performance—*DTU* adopts *accuracy* and *completeness* of MVS reconstruction in mean error distance metrics (mm, lower the better), while the *Tanks and Temples* utilizes the percentage metrics (%, higher the better). To acquire a summary measure of accuracy and completeness, the *DTU* and the *Tanks and Temples* dataset uses the arithmetic mean (termed as *overall score*) and the harmonic mean (termed as *F-score*) of them, respectively. 2) Depth estimation performance—*BlendedMVS* adopts *end point error* (EPE), *1-threshold error* e_1, and *three-threshold error* e_3 to measure the depth estimation quality. Please refer to the Appendix for more detailed illustration on datasets, evaluation metrics, and implementation details.

B. Benchmark Performance

Benchmark on DTU dataset: We benchmark our method on the *DTU evaluation set* and conduct a comprehensive comparison with traditional (geometric) and state-of-the-art learning-based MVS approaches. We follow the standard evaluation procedure [10] for quantitative benchmark and summarize the *mean error distance* metrics (in mm, lower the better) including reconstruction *accuracy*, *completeness*, and *overall score*, as shown in Table I. With different settings, including the changes of N, τ, and N_c (detailed in the Table II of the Appendix), our method performs an excellent tradeoff between the reconstruction *accuracy* and *completeness*. It achieves the best performance in terms of the *accuracy*, *completeness*, and *overall score* compared with the existing traditional and learning-based methods, indicating the state-of-the-art performance of our method. We qualitatively compare the depth estimation and reconstruction results of several reflective and low-textured scenes with illumination changes on *DTU evaluation set* in Figs. 3 and 4, respectively, where our method achieves more complete depth estimation and dense point cloud reconstruction with fine-grained details preserved

TABLE I
QUANTITATIVE BENCHMARKING RESULTS ON *DTU EVALUATION SET* ($W \times H = 1152 \times 864$)

Methods	Year	Mean error distance		
		ACC. ↓ (mm)	Comp. ↓ (mm)	Overall ↓ (mm)
Gipuma [27]	2015	0.283	0.873	0.578
COLMAP [28], [29]	2016	0.400	0.664	0.532
MVSNet [1]	2018	0.396	0.527	0.462
R-MVSNet [2]	2019	0.385	0.459	0.422
AttMVS [17]	2020	0.383	0.329	0.356
CasMVSNet [3]	2020	0.325	0.385	0.355
AA-RMVSNet [5]	2021	0.376	0.339	0.357
EPP-MVSNet [6]	2021	0.413	0.296	0.355
PatchMatchNet [30]	2021	0.427	0.277	0.352
Vis-MVSNet [4]	2022	0.369	0.361	0.365
IterMVS† [31]	2022	0.373	0.354	0.363
IS-MVSNet [32]	2022	0.351	0.359	0.355
BH-RMVSNet [20]	2022	0.368	0.303	0.335
CDS-MVSNet† [7]	2022	0.365	0.281	0.323
UniMVSNet† [8]	2022	0.364	0.279	0.321
TransMVSNet† [9]	2022	0.360	0.271	0.316
NP-CVP-MVSNet [33]	2022	0.356	0.275	0.315
MVSTER† [34]	2022	0.350	0.276	0.313
CostFormer [35]	2023	0.378	0.313	0.345
DSC-MVSNet† [36]	2023	0.316	0.372	0.344
IGEV-MVS† [37]	2023	0.331	0.316	0.324
N2MVSNet [38]	2023	0.336	0.295	0.316
DispMVS [39]	2023	0.354	0.324	0.339
Ours ($N = 5, \tau = 0.3, N_c = 6$)		**0.262**	0.539	0.401
Ours ($N = 5, \tau = 0.3, N_c = 5$)		0.285	0.427	0.356
Ours ($N = 7, \tau = 0.1, N_c = 4$)		0.317	0.323	0.320
Ours ($N = 5, \tau = 0.1, N_c = 3$)		0.368	0.263	0.315
Ours ($N = 7, \tau = 0.1, N_c = 3$)		0.364	**0.262**	**0.313**

↓ The ↓ means that the smaller value indicates the better MVS performance.
† Re-evaluated in the same platform as ours by utilizing the released optimal checkpoints.

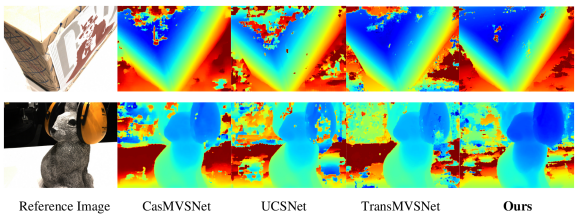


Fig. 3. Qualitative comparison of the depth map estimations on Scan13 (1st row) and Scan33 (2nd row) of the *DTU evaluation set*.

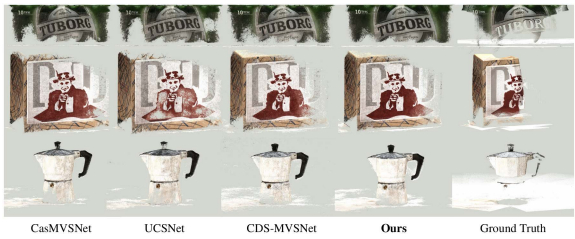


Fig. 4. Qualitative comparison of the point cloud reconstruction of Scan12 (1st row), Scan13 (2nd row), and Scan77 (3rd row) on the *DTU evaluation set*.

benefiting from the proposed LCM scheme, qualitatively verifying the quantitative comparison results.

Benchmark on Tanks and Temples dataset: We benchmark our method on both the *intermediate set* and the *advanced set* of the *Tanks and Temples* benchmark and report the *F-score* (in %, higher the better) in Table II. Our method achieves competitive reconstruction performance compared to the state of the arts, demonstrating the effectiveness and strong generalization ability of our method on both indoor and outdoor scenes. The visualization of the reconstruction errors, as shown in Fig. 5, indicates the superiority of our method in comparison to the state of the arts.

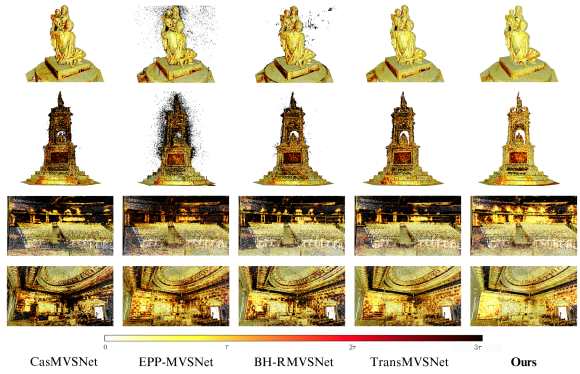


Fig. 5. Visualization of the reconstruction errors of the four scenes including *family*, *francis*, *auditorium*, and *courtroom* on the Tanks and Temples benchmark. τ is the per-scene point distance threshold defined by the benchmark and darker color indicates a larger reconstruction error with respect to τ.

Notably, our method underperforms when reconstructing several scenes with strong backlight or slim structures, we analyze the specific reason in the Appendix.

Benchmark on BlendedMVS dataset: To further demonstrate the superiority of our method for accurate depth map estimation, we quantitatively compare our method with state-of-the-art methods on the *BlendedMVS validation set*. We adopt the original input image resolution 768×576 and set the number of input views to 5 for all methods to ensure a fair comparison. As shown in Table III, our method obtains impressive results with the lowest EPE, e_1, and e_3, showing the ability of our method for inferring high-quality depth maps.

C. Ablation Study

In this section, we conduct systematic ablation experiments to analyze the effectiveness and efficiency of each component of our method. All the ablation experiments are conducted on the *DTU evaluation set*. Please see more ablation studies (BPA, experimental settings) in the Appendix.

Baseline method: CasMVSNet [3] has been serving as the baseline method for almost all the state-of-the-art methods [6], [7], [8], [9] as it proposes the cascade coarse-to-fine cost volume formulation to allow the high-resolution depth map estimation and point cloud reconstruction. To fairly compare with the state of the arts and show the effectiveness of our proposed method components, we adopt CasMVSNet as our baseline method, which applies FPN for feature extraction, the heuristic variance-based scheme for cost volume aggregation, and L_1 loss for optimization.

Bottom–up path augmentation: As aforementioned, we introduce the BPA to augment the propagation of low-level features and incorporate more context information for robust feature matching and continuous depth estimation. Benefiting from BPA, *Model A* addresses the oversmoothing depth estimation (see Fig. 5 of the Appendix) around the object boundary and improves the reconstruction *completeness* (0.370 → 0.344) and *overall score* (0.367 → 0.354).

LCM modules and adapted focal loss: Based on *Model A*, we adopt the proposed *coarse-to-fine LCM* (*Model B*), *efficient LCM*

TABLE II
QUANTITATIVE BENCHMARKING RESULTS ON THE *TANKS AND TEMPLES* BENCHMARK

Methods	Year	*F-score* ↑ (%)															
		Intermediate set									*Advanced Set*						
		Mean	*Fam.*	*Fra.*	*Hor.*	*Lig.*	*M60*	*Pan.*	*Pla.*	*Tra.*	**Mean**	*Aud.*	*Bal.*	*Cou.*	*Mus.*	*Pal.*	*Tem.*
Colmap [28], [29]	2016	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
R-MVSNet [2]	2019	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38	24.91	12.55	29.09	25.06	38.68	19.14	24.96
CasMVSNet [3]	2020	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51	31.12	19.81	38.46	29.10	43.87	27.36	28.11
AttMVS [17]	2021	60.05	73.90	62.58	44.08	**64.88**	56.08	59.39	**63.42**	56.06	31.93	15.96	27.71	37.99	52.01	29.07	28.84
PatchMatchNet [30]	2021	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81	32.31	23.69	37.73	30.04	41.80	28.31	32.29
AA-RMVSNet [5]	2021	61.51	77.77	59.53	51.53	64.02	64.05	59.47	60.85	55.50	33.53	20.96	40.15	32.05	46.01	29.28	32.71
EPP-MVSNet [6]	2021	61.68	77.86	60.54	52.96	62.33	61.69	60.34	62.44	55.30	35.72	21.28	39.74	35.34	49.21	30.00	**38.75**
NP-CVP-MVSNet [33]	2022	59.64	78.93	64.09	51.82	59.42	58.39	55.71	56.07	52.71	-	-	-	-	-	-	-
Vis-MVSNet [4]	2022	60.03	77.40	60.23	47.07	63.44	62.21	57.28	60.54	52.07	33.78	20.79	38.77	32.45	44.20	28.73	37.70
IterMVS [31]	2022	56.94	76.12	55.80	50.53	56.05	57.68	52.62	55.70	50.99	34.17	25.90	38.41	31.16	44.83	29.59	35.15
CDS-MVSNet [7]	2022	61.58	78.85	63.17	53.04	61.34	62.63	59.06	62.28	52.30	-	-	-	-	-	-	-
BH-RMVSNet [20]	2022	61.96	78.62	62.73	51.21	62.13	63.59	61.09	60.85	55.50	34.81	25.79	40.09	34.50	44.89	29.08	34.51
IS-MVSNet [32]	2022	62.82	79.92	62.05	52.54	62.68	63.65	62.57	62.94	56.21	34.87	20.54	39.88	33.07	47.73	30.12	37.91
TransMVSNet [9]	2022	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
DCS-MVSNet [36]	2023	53.48	68.06	47.43	41.60	54.96	56.73	53.86	53.46	51.71	-	-	-	-	-	-	-
CostFormer [35]	2023	57.10	74.22	56.27	54.41	56.65	54.46	51.15	57.65	51.70	34.31	26.77	39.13	31.58	44.55	28.79	35.03
DispMVS [39]	2023	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
N2MVSNet [38]	2023	62.14	80.39	65.64	51.08	62.33	62.30	61.89	59.02	54.47	-	-	-	-	-	-	-
Ours (LCM-MVSNet)		**63.33**	**81.24**	**63.98**	52.56	63.00	**65.44**	**63.13**	61.51	55.76	**38.54**	**27.22**	**44.73**	**39.21**	**53.02**	**32.73**	34.33

↑ The ↑ means that the larger value indicates the better MVS performance.
– The – denotes that the method does not report the MVS performance on the benchmark.

TABLE III
QUANTITATIVE BENCHMARKING RESULTS ON *BLENDEDMVS VALIDATION SET*

Methods	EPE ↓	e_1 ↓ (%)	e_3 ↓ (%)
MVSNet [1]	1.49	21.98	8.32
CVP-MVSNet [40]	1.90	19.73	10.24
CDS-MVSNet [7]	1.80	22.88	9.28
Vis-MVSNet [4]	1.56	21.68	8.36
Cas-MVSNet [3]	1.43	19.73	10.24
EPP-MVSNet [6]	1.17	12.66	6.20
UniMVSNet [8]	1.17	11.27	4.96
TransMVSNet [9]	1.05	13.74	5.47
Ours	**1.02**	**10.15**	**4.54**

The bold values indicate the best performance.

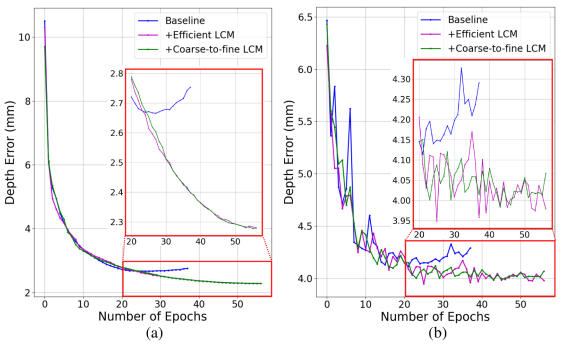


Fig. 6. Comparison of *mean absolute depth error* by proposed LCM modules for cost volume aggregation (a) during training on the *DTU training set*; (b) during validation on the *DTU validation set*.

(*Model C*) to adaptively fuse multi-view feature volumes into the cost volume. Fig. 6(a) and (b) shows the descent curve of the mean absolute depth error by different cost volume aggregation modules on the *DTU training* and *validation set*, respectively, demonstrating that the models with LCM modules produce more accurate depth estimations. The quantitative ablation results on the *DTU evaluation set* shown in Table IV further verify that both LCM modules effectively improve the reconstruction *accuracy*, *completeness*, and *overall score* by a large margin. Notably, the adapted focal loss effectively boosts the reconstruction quality

TABLE IV
ABLATION EXPERIMENTS ON DIFFERENT COMPONENTS OF PROPOSED METHOD ($N = 5$, $W \times H = 1152 \times 864$, $\tau = 0.3$, AND $N_c = 3$)

Models	Feature extraction		Cost volume aggregation				Loss function		Mean error distance ↓ (mm)		
	FPN	FPN+BPA	Variance	AA	C-LCM	E-LCM	L1	Focal	Acc.	Comp.	Overall
Baseline	✓		✓				✓		0.364	0.370	0.367
Model A		✓	✓				✓		0.364	0.344	0.354
Model B		✓			✓		✓		0.355	0.335	0.345
Model C		✓				✓	✓		**0.348**	0.331	0.340
Model D		✓			✓			✓	0.358	0.275	**0.317**
Model E		✓				✓		✓	0.362	**0.274**	0.318
Model G		✓		✓			✓		0.356	0.334	0.345
Model H		✓		✓				✓	0.364	0.279	0.321

The bold values indicate the best performance.

TABLE V
ABLATION EXPERIMENTS ON RUNTIME AND MEMORY

Models	Memory ↓ (MB)	Runtime ↓ (s)	*F-score* ↑ (%)
Baseline	10427	0.901	44.90
+ABN	9115	**0.834**	44.90
+ABN+AA	8119	0.944	40.86
+ABN+C-LCM (Ours)	**7195**	0.862	**45.24**
+ABN+E-LCM (Ours)	**6910**	**0.760**	**45.20**

The bold values indicate the best performance.

in terms of *completeness* and *overall score* to the state-of-the-art performance (*Model D* and *Model E*).

Comparison with adaptive aggregation (AA) module: We compare the proposed LCM modules with AA module [5], [8] (*Model G* and *Model H*, as shown in Table IV) adopting an additional reweight network to compute the weight for each feature volume. The results show that our *coarse-to-fine LCM* (*Model B, D*) and *efficient LCM* (*Model C, E*) outperforms AA in terms of reconstruction *accuracy*, *completeness*, and *overall score*.

Runtime and memory: To further show the efficiency and effectiveness of the proposed LCM modules, we report the memory footprint, inference time per image (runtime), and *F-score*, as shown in Table V. Here, we evaluate the memory footprint on the *DTU training set*, and then test the runtime and *F-score* on the *Tanks and Temples training set* to verify the generalization ability of the proposed modules. For memory footprint, the introduction of ABN reduces memory footprint by 12.58%. The *coarse-to-fine LCM* and *efficient LCM* further

shrink memory footprint by 21.06% and 24.19%, respectively. For runtime, the integration of ABN and *coarse-to-fine LCM* effectively speeds up the runtime by 4.33% while the integration of ABN and *efficient LCM* significantly speeds up the runtime by 15.65%. Compared to AA module, ours are more efficient in terms of memory footprint and runtime, concurrently possessing higher *F-score* indicating stronger generalization ability.

D. Scalability

We demonstrate the scalability of our method in two-fold: domain adaptability and scale adaptability.

For the domain adaptability (outdoor → indoor), we regard it as one of the bases to verify the scalability because the illumination condition, scene texture distribution, depth range, and scene scale greatly vary from outdoor to indoor scenes. Although our method is fine-tuned on the outdoor scenes from the *BlendedMVS training set*, our method achieves superior reconstruction performance over the state of the arts for all the indoor scenes (*auditorium*, *ballroom*, *courtroom*, and *museum*) on the *advanced set* of the *Tanks and Temples* benchmark, as shown in Table II. Specifically, our method outperforms the TransMVSNet [9] (with transformer) by 2.38%, 0.14%, 4.44%, 6.53% in terms of *F-score* on the *auditorium*, *ballroom*, *courtroom*, and *museum*, respectively. We visualize and compare the reconstruction errors (darker color indicates a larger reconstruction error) of our method and state-of-the-art methods on complex indoor scenes *auditorium* and *courtroom*, as shown in Fig. 5, where our method exhibits more accurate and complete reconstruction.

For the scale adaptability (tiny scale → large scale), we naturally regard it as an indication of the scalability because it represents the adaptability of the method to the scene scale. We first differentiate the scene scale [41] as follows.

1) Tiny-scale scene, such as an object on the table.
2) Small-scale scene, such as a statue.
3) Medium-scale scene, such as a single architecture.
4) Large-scale scene, such as a group of architectures.

We then visualize point cloud reconstruction results on multiscale scenes with varying depth ranges. As shown in Fig. 7 (a), our method reconstructs fine-grained details of the tiny-scale objects with highly complex curved structure. Fig. 7(b) shows that our method performs complete dense reconstruction on small-scale scenes with specular and low-textured surfaces. Fig. 7(c) demonstrates that our method can reconstruct medium-scale scenes in high completeness with fine details. Fig. 7(d) shows the effectiveness of our method in generating complete reconstruction on large-scale scenes. In conclusion, our method can scale from tiny-scale scenes to large-scale scenes with competitive reconstruction performance.

IV. Real-World Application for UAV-Based Infrastructure Defect Inspection and Localization

We deploy our method into our UAV-based infrastructure defect inspection framework (see Fig. 1) for infrastructure reconstruction and defect localization, with crack as our target defect. As shown in Fig. 8(a), we adopt three UAVs to cooperatively collect multi-view images for reconstruction and close-range images for defect inspection autonomously (see Appendix for more illustration).

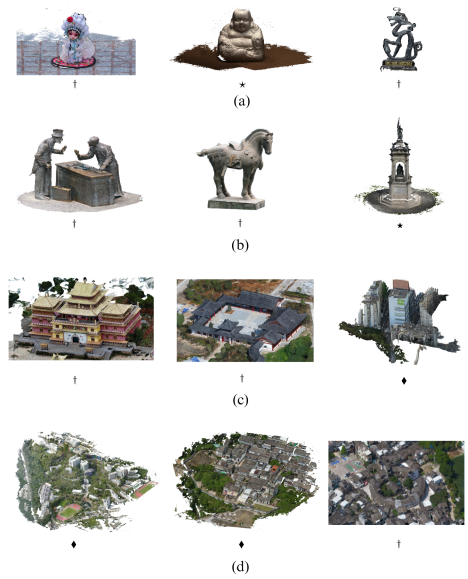


Fig. 7. Point cloud reconstruction results of multiscale scenes with varying depth ranges. Symbol †, ⋆, ★, and ♦ under each reconstruction result denotes that the corresponding input multi-view images are from *BlendedMVS*, *DTU*, *Tanks*, and *Temples*, and our *self-collected* dataset, respectively. (a) Tiny-scale scenes. (b) Small-scale scenes. (c) Medium-scale scenes. (d) Large-scale scenes.

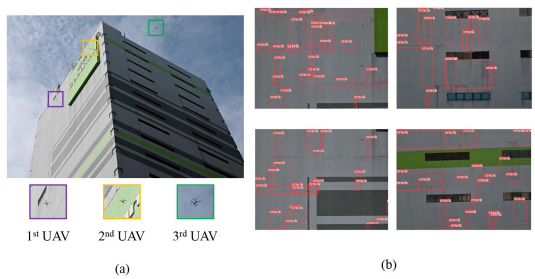


Fig. 8. Experiments for (a) multi-UAV-based data collection; (b) defect detection results.

Reconstruction: 826 multi-view images with the resolution of 1152 × 832 are collected for 3-D reconstruction. As shown in Fig. 1, we follow our MVS method (**green** block) to achieve complete and dense point cloud reconstruction of the target warehouse while preserving highly detailed texture information. Our MVS method outperforms the MVS solutions adopted in the state-of-the-art defect localization methods [23], [42], [43], as shown in Table VII (Appendix) and Fig. 9. In comparison to *DJI TERRA* (6 h 39 mins), our MVS method (44.928 mins: 24.378 mins for SfM, 6.569 mins for view selection, and 13.981 mins for MVS) significantly speeds up the reconstruction process by 8.88 times on an NVIDIA RTX 3090Ti GPU, facilitating the overall defect inspection process (see Appendix for more runtime analysis).

Defect detection and localization: We train YOLOv6-l on our self-established defect dataset (see Appendix) and it achieves

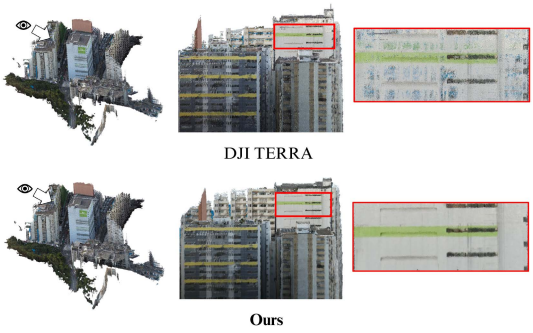


Fig. 9. Comparison of point cloud reconstruction between *DJI TERRA* and our method.

82% mean average precision on the validation set. We then test it on the collected 923 close-range images. As shown in Fig. 8(b), the model detects small cracks with low false and miss detection rates. We adopt the similar methodology as [44] to register the detected cracks onto the reconstructed infrastructure model. The global geographic coordinates of the detected defects can be achieved, providing a solid reference for the maintenance.

V. Limitations

First, our network produces less accurate depth estimation under specular scene with extreme backlight and less complete depth estimation under scene with slim structure. Second, the uneven distribution of sparse points from SfM will affect the computation accuracy of the normalized matching score, which may limit the reconstruction performance of our method. Third, similar to other state of the arts, our method is sensitive to the inference hyperparameters, such as the number of input views, the number of consistent views, and probability threshold.

VI. Conclusion

We presented the LCM-MVSNet for accurate and complete multi-view depth estimation and dense point cloud reconstruction, proposed the LCM scheme for adaptive cost volume aggregation, enhanced the shallow feature information flow to smooth depth estimation, and adapted the focal loss into the end-to-end MVS supervision to reduce ambiguity. The LCM-MVSNet was extensively evaluated on three benchmark datasets to verify the point cloud reconstruction performance and depth estimation quality. The experimental results showed that the proposed LCM-MVSNet achieves 0.313 mm *overall score*, 63.33% mean *F-score*, and 38.54% mean *F-score* on the *DTU evaluation set*, *Tanks and Temples intermediate set*, and *Tanks and Temples advanced set*, respectively, demonstrating the superior point could reconstruction performance. The benchmarking results on the *BlendedMVS validation set* presented the state-of-the-art depth estimation performance of our method with lowest estimation error.

We also made a step toward automated infrastructure inspection by deploying our LCM-MVSNet into our UAV-based infrastructure defect inspection framework for infrastructure reconstruction. In our future work, we will develop depth refinement module to enhance the depth estimation and develop dynamic consistency checking strategy to improve the reconstruction.

References

[1] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 767–783.
[2] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5520–5529.
[3] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2492–2501.
[4] J. Zhang, S. Li, Z. Luo, T. Fang, and Y. Yao, "VIS-MVSNet: Visibility-aware multi-view stereo network," *Int. J. Comput. Vis.*, vol. 131, pp. 199–214, 2022.
[5] Z. Wei, Q. Zhu, C. Min, Y. Chen, and G. Wang, "AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6167–6176.
[6] X. Ma, Y. Gong, Q. Wang, J. Huang, L. Chen, and F. Yu, "EPP-MVSNet: Epipolar-assembling based depth prediction for multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5712–5720.
[7] K. T. Giang, S. Song, and S. Jo, "Curvature-guided dynamic scale networks for multi-view stereo," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–19.
[8] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8635–8644.
[9] Y. Ding et al., "TransMVSNet: Global context-aware multi-view stereo network with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8575–8584.
[10] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 153–168, 2016.
[11] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.
[12] T. Schops et al., "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2538–2547.
[13] Y. Yao et al., "BlendedMVS: A large-scale dataset for generalized multi-view stereo networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1787–1796, doi: 10.1109/CVPR42600.2020.00186.
[14] A. Kendall et al., "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 66–75, doi: 10.1109/ICCV.2017.17.
[15] Q. Xu and W. Tao, "Learning inverse depth regression for multi-view stereo with correlation cost volume," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12508–12515.
[16] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo, "P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10451–10460.
[17] K. Luo, T. Guan, L. Ju, Y. Wang, Z. Chen, and Y. Luo, "Attention-aware multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1587–1596.
[18] H. Yi et al., "Pyramid multi-view stereo net with self-adaptive view aggregation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 766–782.
[19] J. Yan et al., "Dense hybrid recurrent multi-view stereo net with dynamic consistency checking," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 674–689.
[20] Z. Wei, Q. Zhu, C. Min, Y. Chen, and G. Wang, "Bidirectional hybrid LSTM based recurrent neural network for multi-view stereo," *IEEE Trans. Vis. Comput. Graph.*, to be published, doi: 10.1109/TVCG.2022.3165860.
[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.
[22] X. Li et al., "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21002–21012.
[23] K. Liu and B. M. Chen, "Industrial UAV-based unsupervised domain adaptive crack recognitions: From database towards real-site infrastructural inspections," *IEEE Trans. Ind. Electron.*, vol. 70, no. 9, pp. 9410–9420, Sep. 2023, doi: 10.1109/TIE.2022.3204953.
[24] S. R. Bulo, L. Porzi, and P. Kontschieder, "In-place activated batchnorm for memory-optimized training of DNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5639–5647.

[25] S. Cheng et al., "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2521–2531.
[26] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *Int. J. Comput. Vis.*, vol. 120, pp. 153–168, 2016.
[27] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multi-view stereopsis by surface normal diffusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 873–881, doi: 10.1109/ICCV.2015.106.
[28] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4104–4113, doi: 10.1109/CVPR.2016.445.
[29] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 501–518.
[30] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "PatchmatchNet: Learned multi-view patchmatch stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14189–14198.
[31] F. Wang, S. Galliani, C. Vogel, and M. Pollefeys, "IterMVS: Iterative probability estimation for efficient multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8596–8605.
[32] L. Wang, Y. Gong, X. Ma, Q. Wang, K. Zhou, and L. Chen, "IS-MVSNet: Importance sampling-based MVSNet," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 668–683.
[33] J. Yang, J. M. Alvarez, and M. Liu, "Non-parametric depth distribution modelling based depth inference for multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8616–8624.
[34] X. Wang et al., "MVSTER: Epipolar transformer for efficient multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 573–591.
[35] W. Chen et al., "CostFormer: Cost transformer for cost aggregation in multi-view stereo," in *Proc. Int. Joint Conf. Artif. Intell.*, 2023, pp. 599–608.
[36] S. Zhang et al., "DSC-MVSNet: Attention aware cost volume regularization based on depthwise separable convolution for multi-view stereo," *Complex Intell. Syst.*, vol. 9, pp. 6953–6969, 2023.
[37] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21919–21928.
[38] Z. Zhang, H. Gao, Y. Hu, and R. Wang, "N2MVSNet: Non-local neighbors aware multi-view stereo network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.
[39] Q. Yan, Q. Wang, K. Zhao, B. Li, X. Chu, and F. Deng, "Rethinking disparity: A depth range free multi-view stereo based on disparity," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 3091–3099.
[40] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4876–4885.
[41] Y. Furukawa et al., "Multi-view stereo: A tutorial," *Foundations Trends Comput. Graph. Vis.*, vol. 9, no. 1–2, pp. 1–148, 2015.
[42] C. Zhang, M. Jamshidi, C.-C. Chang, X. Liang, Z. Chen, and W. Gui, "Concrete crack quantification using voxel-based reconstruction and Bayesian data fusion," *IEEE Trans. Ind. Informat.*, vol. 18, no. 11, pp. 7512–7524, Nov. 2022, doi: 10.1109/TII.2022.3147814.
[43] G. Winkelmaier, R. Battulwar, M. Khoshdeli, J. Valencia, J. Sattarvand, and B. Parvin, "Topographically guided UAV for identifying tension cracks using image-based analytics in open-pit mines," *IEEE Trans. Ind. Electron.*, vol. 68, no. 6, pp. 5415–5424, Jun. 2021, doi: 10.1109/TIE.2020.2992011.
[44] Y. Tan, G. Li, R. Cai, J. Ma, and M. Wang, "Mapping and modelling defect data from UAV captured images to BIM for building external wall inspection," *Automat. Construction*, vol. 139, 2022, Art. no. 104284.

Guidong Yang (Graduate Student Member, IEEE) received the B.Eng. degree in mechanical engineering (honored class) from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2018, the M.Eng. degree in vehicle engineering from SJTU, in 2021, and the M.Sc. degree in mechanical engineering from the Polytecnic University of Milan, Milan, Italy, in 2021. He is currently working toward the Ph.D. degree in mechanical and automation engineering, The Chinese University of Hong Kong, Shatin, Hong Kong.

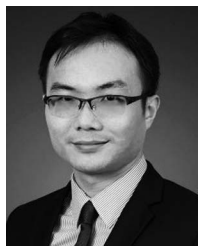
His research interests include 3-D reconstruction and object detection.

Xunkuai Zhou is currently working toward the Ph.D. degree in control science and engineering from Tongji University, Shanghai, China.

He is currently involved in research works as a visiting Ph.D. student with the Department of Mechanical and Automation, The Chinese University of Hong Kong, Shatin, Hong Kong. His research interests include model compression, object detection, object tracking, and 3-D reconstruction.

Chuanxiang Gao received the B.Eng. degree from Northwestern Polytechnical University, Xian, China, in 2020. He is currently working toward the Ph.D. degree in mechanical and automation engineering, The Chinese University of Hong Kong, Shatin, Hong Kong.

His research interests include task planning and motion planning.



Xi Chen received the B.S. Degree in energy and environment systems engineering from Zhejiang University, in 2009, the M.Phil and Ph.D. degrees in building services engineering from the Hong Kong Polytechnic University, in 2011 and 2017 respectively.

He is currently a Research Assistant Professor with The Chinese University of Hong Kong, Shatin, Hong Kong. He has more than ten year experience in sustainable building technology related to the urban energy systems, renewable application in buildings and built environment modeling, and has led or managed multiple research projects including ARC, MOST, RGC, and consultancy projects with the local government and industry. He has authored or coauthored more than 40 papers in peer-reviewed international journals and coauthored a book in green building and renewable application areas.

Dr. Chen was a recipient of the DECRA Fellow in the Australian Research Council and Fulbright Scholar in the Lawrence Berkeley National Laboratory. In addition, he is the Editorial Board Member of Buildings, Energies and Advances in Applied Energy.

Ben M. Chen (Fellow, IEEE) received the B.S. degree in mathematics and computer science from Xiamen University, Xiamen, China, in 1983, the M.S. degree in electrical engineering from Gonzaga University, Spokane, WA, USA, in 1988, and the Ph.D. degree in electrical and computer engineering from Washington State University, Pullman, WA, USA, in 1991.

He is currently a Professor of mechanical and automation engineering with The Chinese University of Hong Kong, Shatin, Hong Kong. From 1992 to 1993, he was an Assistant Professor with the Department of Electrical Engineering, State University of New York, Stony Brook, NY, USA. In 2018, he was a Provost's Chair Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. He has authored or coauthored hundreds of journal and conference articles, and a dozen research monographs in control theory and applications, unmanned systems, and financial market modeling. His research focuses on unmanned systems and their applications.

Dr. Chen is a Fellow of the Academy of Engineering, Singapore. He was on the editorial boards of a dozen international journals including *Automatica* and IEEE TRANSACTIONS ON AUTOMATIC CONTROL. He is currently an Editor-in-Chief for *Unmanned Systems*, and the Editor for the *International Journal of Robust and Nonlinear Control*.