A Novel Street View MVS Pipeline with Edge-Enhanced Sky Masking and Cross Algorithm Data Fusion

Zequan Chen [1,2], Dong Xu [1,3], Ming Zhang [1,2], Yang Liu [1,2], Zexin Yang [1,3], Keke Liu [1,3] and Chang Sun [4]

[1] Guangzhou Urban Planning & Design Survey Research Institute Co., Ltd, Guangzhou 510060, China - (zequanchen, dongxu, ming.zhang, liuyang, zexinyang)@gzpi.com.cn
[2] Guangdong Enterprise Key Laboratory for Urban Sensing, Monitoring and Early Warning, Guangzhou 510060, China
[3] Collaborative Innovation Center for Natural Resources Planning and Marine Technology of Guangzhou, Guangzhou 510060, China
[4] School of Civil Engineering, Central South University, Changsha 410075, China - 8210222102@csu.edu.cn

Keywords: MVS; street view; data fusion; point cloud; reconstruction.

Abstract

MVS (Multi-View Stereo) establishes dense correspondences among multiple calibrated images to generate 3D point clouds, which has broad applications in fields such as 3D modeling, robot navigation, and autonomous driving. In street view MVS, the distant, weakly-textured sky pixels in images significantly degrade the quality of the generated point cloud, manifesting as pronounced edge noise at building boundaries, and the completeness of the point cloud requires further improvement. Therefore, we design an Edge-enhanced Sky Masking Module to free street view MVS from sky interference, reducing edge noise by approximately 40%. In addition, we propose a Fusion Module based on Local Planarity Features, which integrates the strengths of both traditional and learning-based algorithms to generate superior dense point clouds, outperforming current mainstream methods in terms of completeness and F1 score.

1. Introduction

Multi-View Stereo (MVS) represents a core research area of computer vision, aimed at reconstructing dense geometric representations of real-world scenes from multiple overlapping photographs (Su and Tao, 2025). This capability demonstrates significant value with diverse applications across various fields, including autonomous driving, robotic navigation, augmented reality and virtual reality (AR/VR), cultural heritage reconstruction, and 3D modelling, among others. (Cao et al., 2021; Gao, 2024; Ning et al., 2025; Shao et al., 2025; Song et al., 2022)

In open street-view reconstruction, distant, weakly-textured sky regions present substantial challenges, causing inaccurate depth estimations at building-sky boundaries and generating pronounced edge noise (Rich et al., 2025) in reconstructed dense point clouds. Consequently, effective reconstruction of complex street-view scenes remains a critical research objective.

Following decades of advancement, MVS methodology has matured, incorporating key processes including camera calibration, feature extraction/matching, Structure from Motion (SfM) (Schonberger and Frahm, 2016), stereo matching. Traditional MVS relies on accurate feature matching, but exhibits limitations in large-scale scenes and areas with repetitive or weak textures (Yao et al., 2020). With the rapid development of deep learning, learning-based MVS methods have received growing attention (Wang et al., 2025). These methods, leveraging Convolutional Neural Networks (CNNs) and other deep learning modules, have significantly improved the accuracy and efficiency of depth information extraction from images, but still face challenges in complex scenes.

Empirical observations indicate that traditional MVS algorithms are more robust in recovering scene structural information, while learning-based MVS algorithms can better recover weakly-textured areas in scenes. Therefore, we propose a novel street view MVS pipeline that frees the noise components in the sky area and leverages the advantages of both traditional and learning-based methods. The specific contributions include:

1. An Edge-enhanced Sky Masking Module to eliminate weakly-textured interference from sky while preserving objects of interest, effectively reducing edge noise at buildings and constructing cleaner dense point clouds;
2. A Fusion Module based on Local Planarity Features, which fuses the advantageous structures of traditional and learning-based methods to generate more complete dense point clouds.

2. Related Work

Traditional MVS algorithms are broadly classified into three types: volumetric (Ulusoy et al., 2017), point cloud (Furukawa and Ponce, 2010), and depth map (Kar et al., 2017). Of these, depth map-based methods estimate depth maps for individual images through patch matching with photometric consistency, and then fuse these depth maps into dense 3D representations. This pipeline decomposes the reconstruction problem into depth estimation per viewpoint and depth maps fusion, significantly enhancing flexibility and scalability. Gipuma (Galliani et al., 2015), building upon the PatchMatch algorithm (Bleyer et al., 2011), aggregates image similarity across multiple views to obtain more accurate depth maps, while incorporating enhancements to the propagation scheme to enable large-scale parallelization on common GPUs. Colmap (Schönberger et al., 2016) is a representative work with contributions in multiple aspects, including the joint estimation of depth and normal

information, pixelwise view selection using photometric and geometric priors, reliable depth/normal filtering and fusion, etc. Recent work by Xu et al. has significantly advanced such algorithms. ACMH (Xu and Tao, 2019) employs Adaptive Checkerboard sampling and Multi-Hypothesis joint view selection infer the aggregation view subset at each pixel, combined with multi-scale geometric consistency guidance to enhance achieve depth estimation. ACMP (Xu and Tao, 2020) incorporates utilize a probabilistic graphical model to embed planar models into PatchMatch MVS, and proposes a multi-view aggregation matching cost that considers both photometric consistency and planar compatibility to obtain more complete dense point clouds. ACMMP (Xu et al., 2022) further designs a MVS guided by multi-scale geometric consistency and assisted by planar priors, enhancing the distinction of blurred areas and improving the algorithm's depth perception capabilities based on previous work.

Learning-based MVS methods have achieved substantial progress in recent years (Wang et al., 2024), which can be categorized as voxel-based methods (Sun et al., 2021), depth map-based methods (Li et al., 2024), NeRF-based methods (P. Wang et al., 2021), and 3DGS-based methods (Guédon and Lepetit, 2024). Notably, depth map-based methods inherit the advantages of traditional methods, decomposing 3D reconstruction into depth estimation and dense fusion. MVSNet (Yao et al., 2018) proposes an end-to-end deep learning architecture that extracts deep visual features from images using CNNs and then performs joint cost optimization for multi-view to achieve fast dense point cloud generation. MVDepthNet (Wang and Shen, 2018) encodes multi-view observations into a cost volume, combines it with the reference image, and uses an encoder-decoder network to estimate depth maps. It is an online method that can continuously estimate depth for a single moving camera. PatchmatchNet (F. Wang et al., 2021) first introduces iterative multi-scale PatchMatch in an end-to-end trainable architecture, proposing a novel, learnable adaptive propagation and evaluation scheme to achieve efficient MVS. Iter-MVS (Wang et al., 2022) proposes a new GRU-based estimator that encodes the per-pixel probability distribution of depth in its hidden state. TransMVSNet (Ding et al., 2022) attempts to apply Transformers to the MVS task, proposing a Feature Matching Transformer that uses intra-image (self) and inter-image (cross) attention to aggregate long-range contextual information within and across images. Wang et al. (Wang et al., 2023) innovatively transplant the deformable convolution idea from deep learning into the traditional PatchMatch-based method, adaptively deforming patches centered on unreliable pixels to expand the receptive field until enough relevant reliable pixels are covered as anchors.

In summary, traditional MVS methods still have advantages in specific scenarios but remain challenged by low-texture regions. Learning-based MVS methods, incorporating novel network architectures and training strategies, have significantly improved performance and generalization capabilities, but still require further optimization to handle complex environments (Luo et al., 2024). We combine the strengths of traditional and learning-based MVS algorithms, obtaining more complete dense point clouds via plane extraction and fusion. Additionally, masking sky pixels from street view images reduces non-target element interference, substantially reducing the proportion of edge noise.

3. Methodology

To mitigate the impact of sky pixels on outdoor scenes and enhance the completeness of dense point clouds, we introduce a novel MVS framework leveraging cross algorithm data fusion, as illustrated in Figure 1.

Firstly, the Edge-enhanced Sky Masking Module is applied to identify and mask sky pixels from the input images. Specifically, following edge detection, extraction, and enhancement, edge maps are obtained. These edge maps are then directly overlaid onto the input images and undergo semantic segmentation to obtain sky masks. The masked images are then fed into the Fusion Module based on Local Planarity Features. Within this module, the masked images are used for both learning-based and traditional MVS. Planar points are extracted from the learning-based MVS point cloud based on their planarity. These planar points are then fused with the denoised traditional MVS point cloud based on distance, yielding the final dense point cloud.

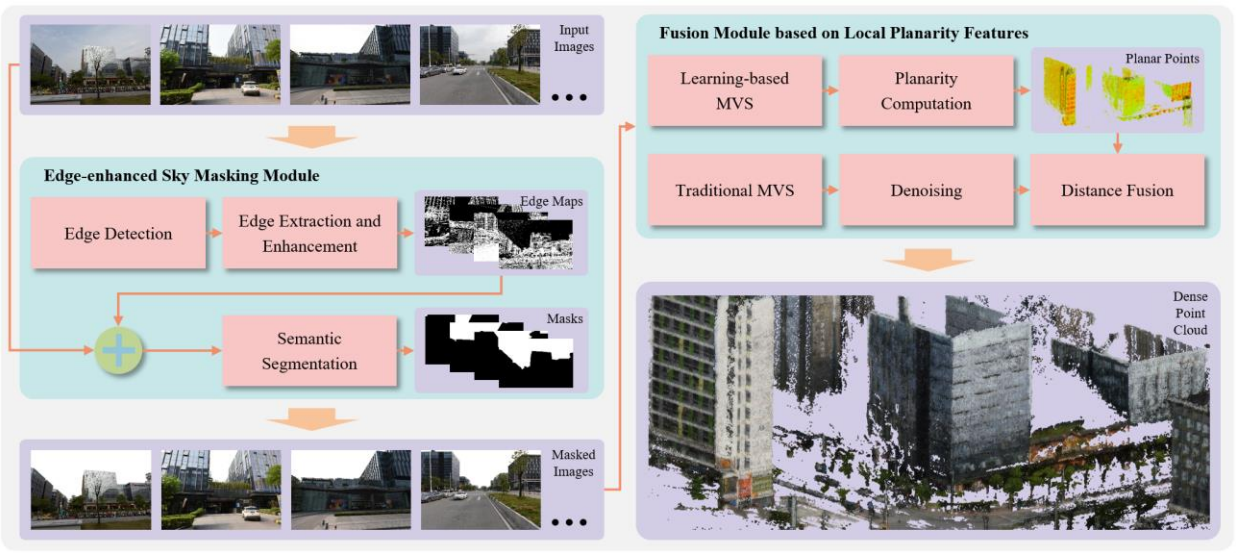

Figure 1. The proposed MVS pipeline.

3.1 Edge-enhanced Sky Masking Module

In outdoor MVS tasks, the sky pixels in the image represent significant long-range, weakly-textured interference, which can degrade the quality of the generated point cloud. To address this issue, we develop an edge-enhanced semantic segmentation module that effectively mask sky pixels while preserving the integrity of objects of interest.

Semantic segmentation is an important branch of image processing and computer vision, designed to interpret image content precisely. It involves classifying pixels and grouping them into the same category. After years of research, image semantic segmentation has advanced significantly and can effectively classify pixels in images. We utilize SegFormer (Xie et al., 2021) for image semantic segmentation, which is capable of distinguishing various types of objects in images, such as buildings, roads, vehicles, pedestrians, sky, vegetation, traffic signs, and signal lights.

During semantic segmentation, pixels of objects adjacent to the sky are prone to being erroneously classified as sky, which can adversely affect subsequent multi-view geometric consistency. To address this problem, we introduce a preprocessing stage that involves edge detection and enhancement to preserve boundary integrity while effectively masking sky pixels. Specifically, the Canny algorithm is employed to detect edges in input images I_{input}. The Canny algorithm utilizes a high threshold T_H and a low threshold T_L to identify strong and weak edges:

$$\begin{aligned} strong\ \ edges &= \{pixels \mid M > T_H\} \\ weak\ \ edges &= \{pixels \mid T_L \le M \le T_H\} \end{aligned} \tag{1}$$

where M is the gradient magnitude:

$$M = \sqrt{(G_x * I)^2 + (G_y * I)^2} \tag{2}$$

where $G_x * I$ and $G_y * I$ are the convolution results of the image with the Sobel operator in the horizontal and vertical directions, respectively:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{3}$$

After edge detection, a set of grayscale images I_{edge} containing edge information is obtained. To enhance the visibility of edges, we apply the dilation operation in mathematical morphology to thicken the edge information in I_{edge}, thereby generating Edge Maps. Edges within the edge maps, represented by black pixels, are superimposed on I_{input} to produce an edge-enhanced image set $I_{enhanced}$.

Finally, semantic segmentation is performed on $I_{enhanced}$, and based on the segmentation results, the sky pixels in I_{input} are masked to generate a masked image set I_{masked} for subsequent steps.

3.2 Fusion Module based on Local Planarity Features

SfM is initially applied on I_{masked}, followed by MVS. Two dense point clouds, C_{trad} and C_{dl}, are generated using traditional and learning-based MVS algorithms, respectively.

Learning-based MVS generally achieves better recovery in weakly-textured areas, such as walls and windows. Therefore,

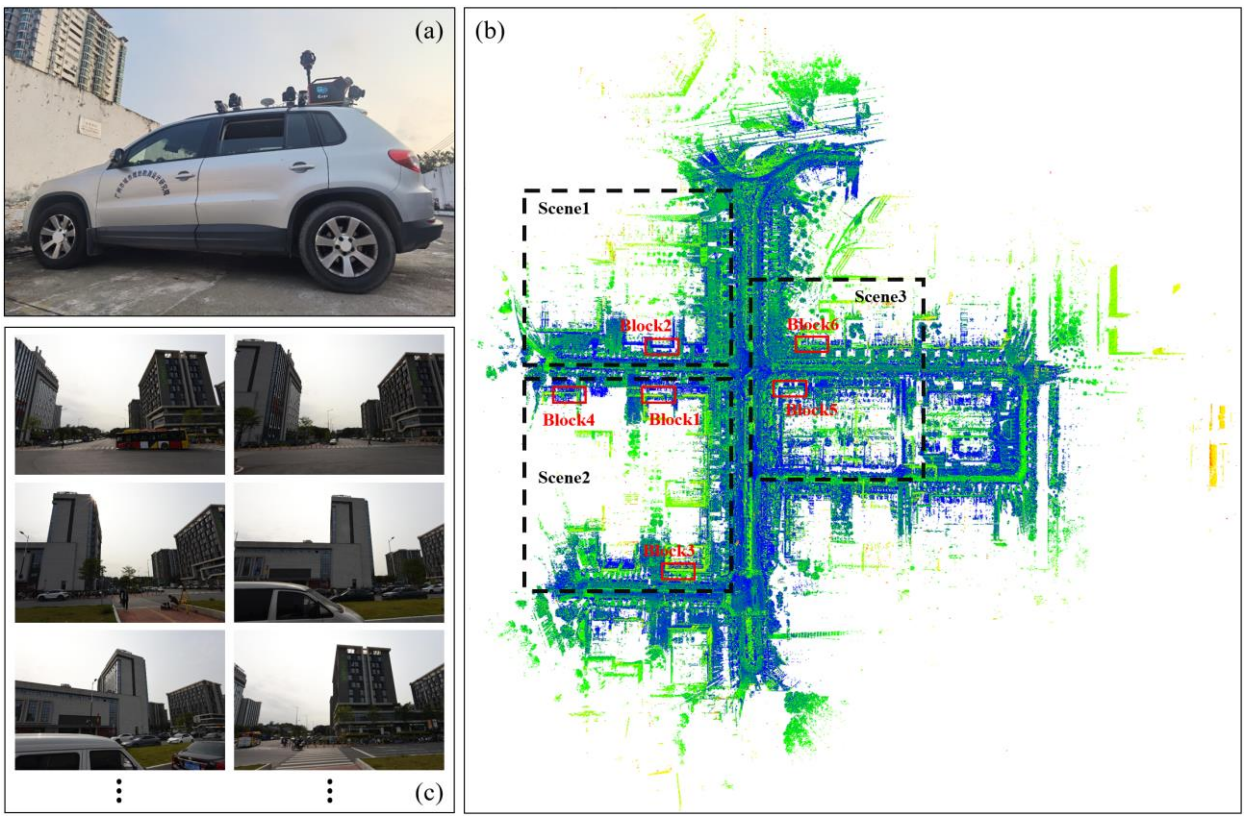

Figure 2. (a) Mobile mapping vehicle, (b) LiDAR point cloud of experimental area, (c) Partial image data.

$$Cov = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)^2 & \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y) & \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(z_i-\mu_z) \\ \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y) & \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_y)^2 & \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_y)(z_i-\mu_z) \\ \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(z_i-\mu_z) & \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_y)(z_i-\mu_z) & \frac{1}{n}\sum_{i=1}^{n}(z_i-\mu_z)^2 \end{bmatrix} \quad (5)$$

local planar features for C_{dl} are calculated first. Specifically, for each point in C_{dl}, a local neighbourhood is defined, and all 3D points within this neighbourhood are denoted as the set $P = \{p_1, p_2, \ldots, p_n\}$, where n is the number of points in the set, and the coordinate of p_i is (x_i, y_i, z_i). Then, the mean coordinates of all points in P are calculated in the x, y, and z directions:

$$\mu_x = \frac{1}{n}\sum_{i=0}^{n} x_i, \mu_y = \frac{1}{n}\sum_{i=0}^{n} y_i, \mu_z = \frac{1}{n}\sum_{i=0}^{n} z_i \quad (4)$$

Based on the mean coordinates calculated above, the covariance matrix Cov is constructed, as shown in equation 5.

Eigenvalue decomposition of the covariance matrix Cov yields three eigenvalues $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$, satisfying $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$. The planarity is then calculated as:

$$P_\lambda = \frac{\lambda_2 - \lambda_3}{\lambda_1} \quad (6)$$

For a planar feature, λ_3 should be significantly smaller than the other two eigenvalues. Therefore, a larger P_λ indicates higher planarity of the point. Given a threshold T_P, points in C_{dl} with $P_\lambda > T_P$ are identified as planar points, which collectively form the planar point cloud C_{planar}.

During the fusion process, noise points can significantly affect the distance calculation between two point clouds. Therefore, we first apply denoising to C_{trad} by removing outliers that are distant from surrounding points. Given a distance threshold T_{D1} and an inlier number threshold T_N, for each point in C_{trad}, if fewer than T_N points are within distance T_{D1}, the point is considered an outlier and is removed. The resulting denoised point cloud is denoted as $C_{denoised}$.

Finally, the distance from each point in C_{planar} to the nearest point in $C_{denoised}$ is calculated. If the distance exceeds the threshold T_{D2}, the point is retained; otherwise, it is removed. Generally, a more complete fusion result can be obtained when $T_{D2} < T_{D1}$. The retained planar points are subsequently integrated with $C_{denoised}$ to produce the final dense point cloud.

4. Experiments and Analysis

4.1 Dataset

The experimental data were collected using well-calibrated LiDAR and multi-view camera equipment mounted atop a mobile mapping vehicle, as shown in Figure 2(a). The experimental area is a typical urban campus scene, covering an area of approximately 700 meters in length and width. The LiDAR point cloud data are presented in Figure 2(b), and approximately 1500 images were captured, as shown in Figure 2(c).

4.2 Results of Sky Masking

The results of directly masking the sky via semantic segmentation and using the edge-enhanced sky masking module described in Section 3.1 are shown in Figure 3. Qualitative analysis reveals that conventional image semantic segmentation algorithms produce less precise separation results at the edges, leading to noticeable jagged errors. In contrast, by enhancing edges prior before semantic segmentation, our method effectively improves the accuracy of edge separation and ensures that sky masking does not degrade the quality of other image elements.

Figure 3. Comparison results of directly sky masking (green rectangle) and edge-enhanced sky masking (red rectangle).

	Scene1	Scene2	Scene3	Avg
Colmap (Schönberger et al., 2016)	16.45	36.26	30.38	27.70
ACMMP (Xu et al., 2022)	86.36	65.89	31.46	61.24
TransMVS (Ding et al., 2022)	29.90	39.17	23.99	31.02
Avg	44.24	47.11	28.61	39.99

Table 1. The proportion (%) of reduced edge noise before and after sky masking.

Three main regions (black dashed rectangles in Figure 2(b)) were captured from the experimental area, and three different algorithms were used for MVS of the images before and after sky masking. The reduction ratio of the number of edge noise outside the building contour was compared, as shown in Table 1. From the table, it can be seen that after preprocessing the images with the proposed edge-enhanced sky masking module, the proportion of edge noise in the generated dense point clouds

is reduced by 39.99% on average, especially for the ACMMP method, where the average reduction in edge noise is as high as 61.24%. These findings confirm that masking the sky from the images effectively reduces the number of edge noise points and improves the cleanliness of the reconstruction results.

4.3 Results of MVS

Six representative blocks (red rectangles in Figure 2(b)) were selected, and the experimental results were evaluated using the evaluation protocol defined in (Schops et al., 2017), including accuracy, completeness, and F1 score at two distance metrics (5cm and 15cm), as shown in Table 2. The visualization result of Block1 is shown in Figure 4. Here, ACMMP and TransMVS were used to generate C_{trad} and C_{dl}, respectively. Our method is not limited to specific algorithm, and theoretically, any traditional and learning-based algorithm can be integrated into our pipeline.

	Method	5cm	15cm
Block1	Colmap	89.00/19.86/32.47	97.32/46.62/63.05
	ACMMP	**98.75**/69.04/81.26	**99.97**/84.95/91.85
	TransMVS	83.68/61.71/71.03	99.16/85.41/91.77
	Ours	91.11/**86.30**/**88.64**	99.56/**98.67**/**99.11**
Block2	Colmap	89.47/21.25/34.34	99.35/63.45/77.44
	ACMMP	**97.92**/58.59/73.31	**99.99**/86.27/92.63
	TransMVS	95.15/89.22/92.09	99.93/98.56/99.24
	Ours	95.66/**93.29**/**94.46**	99.95/**99.84**/**99.90**
Block3	Colmap	83.58/38.96/53.14	97.28/80.21/87.92
	ACMMP	**96.67**/87.77/**92.00**	**99.88**/96.49/98.16
	TransMVS	75.59/59.24/66.42	98.00/88.71/93.13
	Ours	89.46/**92.73**/91.07	99.24/**98.56**/**98.90**
Block4	Colmap	92.95/47.96/63.28	99.74/74.41/85.23
	ACMMP	**98.85**/89.55/93.97	**99.99**/96.37/98.15
	TransMVS	89.54/76.97/82.78	99.88/88.14/93.65
	Ours	94.37/**94.33**/**94.35**	99.94/**98.51**/**99.22**
Block5	Colmap	61.56/23.66/34.18	88.46/75.43/81.43
	ACMMP	**83.66**/46.68/59.93	**99.18**/93.50/96.25
	TransMVS	59.59/**74.16**/66.08	93.80/96.92/95.34
	Ours	79.52/59.73/**68.22**	98.51/**97.94**/**98.23**
Block6	Colmap	40.31/4.84/8.64	85.76/13.47/23.29
	ACMMP	**52.72**/30.90/38.96	94.61/62.61/75.36
	TransMVS	39.30/20.28/26.75	91.39/49.05/63.84
	Ours	51.80/**37.98**/**43.83**	**95.36**/**70.75**/**81.23**

Table 2. Accuracy, completeness, and F1 score (%) of dense point clouds under different metrics (5cm and 15cm). We highlight the **best** and second-best ones in each category.

Quantitative results in Table 2 demonstrate that the accuracy, completeness, and F1 score of our method are consistently achieve top-two performance across all blocks and metrics. The proposed method demonstrates superior performance in terms of completeness and F1 score compared to the other methods, indicating that the proposed method generates more structurally complete dense point clouds while maintaining high accuracy. This is also visually evident from Figure 4, where the dense point cloud generated by our method is more complete than those of the comparison algorithms.

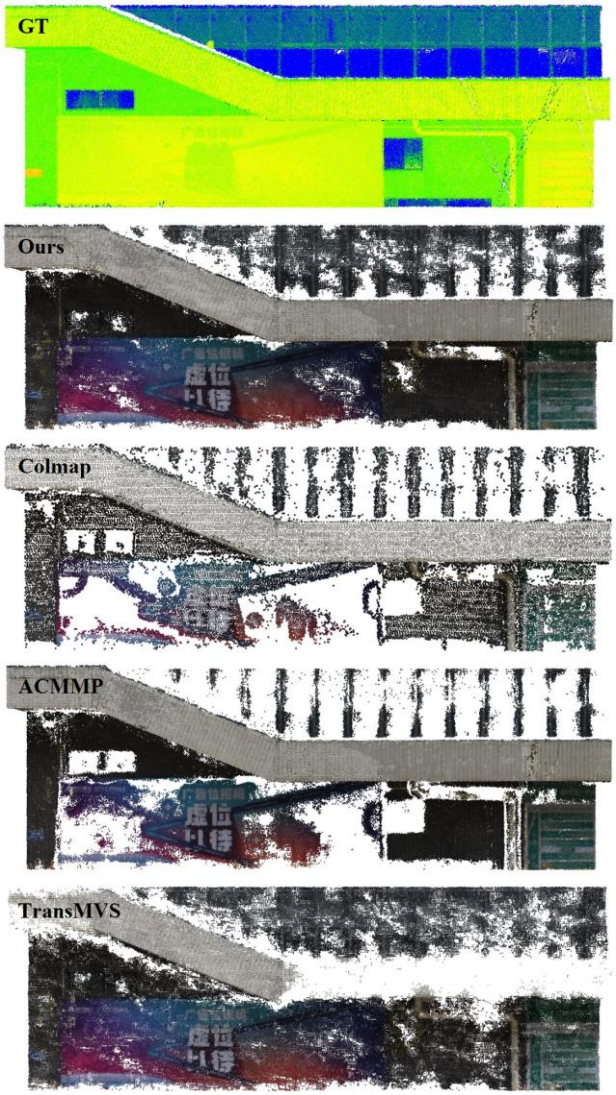

Figure 4. Comparison of dense point clouds of block 1.

The cumulative distribution function (CDF) uniquely characterizes the probability distribution of a random variable. Figure 5 statistically compares the CDFs of accuracy and completeness metrics across all evaluated blocks. Analysis reveals that the ACMMP algorithm typically achieves the best results in accuracy, indicating that the generated point clouds contain less noise. In the contrast, our method outperform all other algorithms in the completeness diagram, confirming its capability to generate more complete dense point clouds once again. Overall, our method achieves comprehensive F1 scores of 86.85% and 97.90% for all blocks under the 5cm and 15cm metrics, respectively, both of which are better than ACMMP's 83.56% and 95.38%, indicating that the dense point clouds generated by our method have superior overall quality.

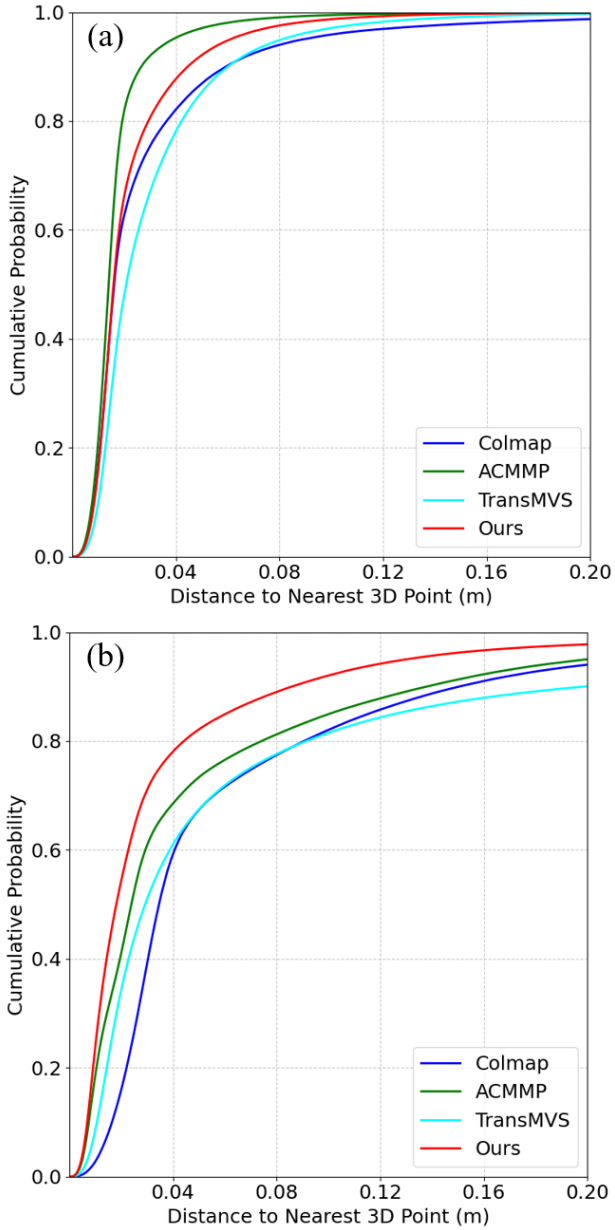

Figure 5. Distribution diagram of (a) accuracy and (b) completeness.

5. Conclusions

Empirical findings demonstrate that traditional MVS methods excel in reconstructing texture-rich structural areas, while learning-based MVS methods excel in reconstructing weakly-textured areas, demonstrating complementary characteristics. Therefore, a fusion method was developed to integrate the strengths of both algorithms by combining the planar parts of learning-based MVS point clouds with traditional MVS point clouds through planarity assessment, resulting in more complete dense point clouds. Additionally, to eliminate sky interference in open street scenes, an edge-enhanced sky masking module was designed to effectively mask sky pixels while preserving regions of interest. Experimental evaluations confirm that our framework generates higher-fidelity dense point clouds against existing mainstream algorithms. In the future, we will explore deeper integration of traditional and learning-based MVS advantages while preserving accuracy, to further improve the quality of street-view dense point clouds.

Acknowledgements

The completion of this work was supported by the National Key Research and Development Program of China under Grant 2022YFB3904105. Thanks for the support of Academic Specialty Group for Urban Sensing in Chinese Society of Urban Planing.

References

Bleyer, M., Rhemann, C., Rother, C., 2011. PatchMatch Stereo - Stereo Matching with Slanted Support Windows, in: *Procedings of the British Machine Vision Conference 2011*, https://doi.org/10.5244/C.25.14.

Cao, M., Zheng, L., Jia, W., Lu, H., Liu, X., 2021. Accurate 3-D Reconstruction Under IoT Environments and Its Applications to Augmented Reality. *IEEE Trans. Ind. Inf*, 17, 2090–2100. https://doi.org/10.1109/TII.2020.3016393.

Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, Xiangyue, Wang, Y., Liu, Xiao, 2022. TransMVSNet: Global Context-aware Multi-view Stereo Network with Transformers, in: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8575–8584. https://doi.org/10.1109/CVPR52688.2022.00839.

Furukawa, Y., Ponce, J., 2010. Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32, 1362–1376. https://doi.org/10.1109/TPAMI.2009.161.

Galliani, S., Lasinger, K., Schindler, K., 2015. Massively Parallel Multiview Stereopsis by Surface Normal Diffusion, in: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 873–881. https://doi.org/10.1109/ICCV.2015.106.

Gao, M., 2024. Research on Cultural Relic Restoration and Digital Presentation Based on 3D Reconstruction MVS Algorithm: A Case Study of Mogao Grottoes' Cave 285, in: *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, pp. 744–754. https://doi.org/10.2991/978-94-6463-540-9_76.

Guédon, A., Lepetit, V., 2024. SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5354-5363.

Kar, A., Häne, C., Malik, J., 2017. Learning a Multi-View Stereo Machine. *Advances in neural information processing systems*, 30.

Li, H., Guo, Y., Zheng, X., Xiong, H., 2024. Learning Deformable Hypothesis Sampling for Accurate PatchMatch Multi-View Stereo. in: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38, No. 4, pp. 3082-3090, https://doi.org/10.1609/aaai.v38i4.28091.

Luo, H., Zhang, J., Liu, X., Zhang, L., Liu, J., 2024. Large-Scale 3D Reconstruction from Multi-View Imagery: A

Comprehensive Review. *Remote Sensing*, 16, 773. https://doi.org/10.3390/rs16050773.

Ning, M., Khajepour, A., Hashemi, E., Sun, C., 2025. A Novel Motion Planning for Autonomous Vehicles Using Point Cloud Based Potential Field. *IEEE Trans. Veh. Technol*, 74, 3780–3792. https://doi.org/10.1109/TVT.2024.3485511.

Rich, A., Stier, N., Sen, P., Höllerer, T., 2025. Smoothness, Synthesis, and Sampling: Re-thinking Unsupervised Multi-view Stereo with DIV Loss, in: *European Conference on Computer Vision (ECCV)*, pp. 380–397. https://doi.org/10.1007/978-3-031-73036-8_22.

Schonberger, J.L., Frahm, J.-M., 2016. Structure-from-Motion Revisited, in: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113. https://doi.org/10.1109/CVPR.2016.445.

Schönberger, J.L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise View Selection for Unstructured Multi-View Stereo, in: *European Conference on Computer Vision (ECCV)*, pp. 501–518. https://doi.org/10.1007/978-3-319-46487-9_31.

Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A., 2017. A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2538–2547. https://doi.org/10.1109/CVPR.2017.272.

Shao, Y., Wang, Q., Sun, H., Ding, X., 2025. Irregular seeds DEM parameters prediction based on 3D point cloud and GA-BP-GA optimization. *Scientific Reports*, 15, 304. https://doi.org/10.1038/s41598-024-84375-3.

Song, S., Kim, D., Choi, S., 2022. View Path Planning via Online Multiview Stereo for 3-D Modeling of Large-Scale Structures. *IEEE Trans. Robot*, 38, 372–390. https://doi.org/10.1109/TRO.2021.3083197.

Su, W., Tao, W., 2025. Context-Aware Multi-view Stereo Network for Efficient Edge-Preserving Depth Estimation. *Int J Comput Vis*, https://doi.org/10.1007/s11263-024-02337-8.

Sun, J., Xie, Y., Chen, L., Zhou, X., Bao, H., 2021. NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video, in: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15593–15602. https://doi.org/10.1109/CVPR46437.2021.01534.

Ulusoy, A.O., Black, M.J., Geiger, A., 2017. Semantic Multi-view Stereo: Jointly Estimating Objects and Voxels, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4531–4540. https://doi.org/10.1109/CVPR.2017.482.

Wang, F., Galliani, S., Vogel, C., Pollefeys, M., 2022. IterMVS: Iterative Probability Estimation for Efficient Multi-View Stereo, in: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8596–8605. https://doi.org/10.1109/CVPR52688.2022.00841.

Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M., 2021. PatchmatchNet: Learned Multi-View Patchmatch Stereo, in: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14189–14198. https://doi.org/10.1109/CVPR46437.2021.01397.

Wang, F., Zhu, Q., Chang, D., Gao, Q., Han, J., Zhang, T., Hartley, R., Pollefeys, M., 2024. Learning-based Multi-View Stereo: A Survey. *arXiv preprint*, arXiv:2408.15235. https://doi.org/10.48550/arXiv.2408.15235.

Wang, K., Shen, S., 2018. MVDepthNet: Real-time Multiview Depth Estimation Neural Network. in: *2018 International conference on 3d vision (3DV)*, pp. 248-257.

Wang, L., She, J., Qiang, Z., Wen, X., Guan, Y., 2025. Transformer-guided Feature Pyramid Network for Multi-View Stereo. *Neurocomputing*, 617, 129066. https://doi.org/10.1016/j.neucom.2024.129066.

Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W., 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *arXiv preprint*, arXiv:2106.10689, https://doi.org/10.48550/ARXIV.2106.10689.

Wang, Y., Zeng, Z., Guan, T., Yang, W., Chen, Z., Liu, W., Xu, L., Luo, Y., 2023. Adaptive Patch Deformation for Textureless-Resilient Multi-View Stereo, in: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1621–1630. https://doi.org/10.1109/CVPR52729.2023.00162.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Advances in neural information processing systems*, 34, 12077-12090, https://doi.org/10.48550/arXiv.2105.15203.

Xu, Q., Kong, W., Tao, W., Pollefeys, M., 2022. Multi-Scale Geometric Consistency Guided and Planar Prior Assisted Multi-View Stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–18. https://doi.org/10.1109/TPAMI.2022.3200074.

Xu, Q., Tao, W., 2019. Multi-Scale Geometric Consistency Guided Multi-View Stereo, in: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5478–5487. https://doi.org/10.1109/CVPR.2019.00563.

Xu, Q., Tao, W., 2020. Planar Prior Assisted PatchMatch Multi-View Stereo. in: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, No. 07, pp. 12516-12523.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. MVSNet: Depth Inference for Unstructured Multi-view Stereo, in: *Proceedings of the European conference on computer vision (ECCV)*, pp. 767-783. https://doi.org/10.1007/978-3-030-01237-3_47.

Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L., 2020. BlendedMVS: A Large-Scale Dataset for Generalized Multi-View Stereo Networks, in: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1787–1796. https://doi.org/10.1109/CVPR42600.2020.00186.