MVSNet: Depth Inference for Unstructured Multi-view Stereo

Dheeraj Chilukuri
MS in Robotics and Autonomous Systems (Mechanical and Aerospace Engineering)
Arizona State University
nchiluk6@asu.edu

Meenakshisundram Ganapathi Subramanian
MS in Robotics and Autonomous Systems (Mechanical and Aerospace Engineering)
Arizona State University
mganapa3@asu.edu

Soorya Boopal
MS in Robotics and Autonomous Systems (Mechanical and Aerospace Engineering)
Arizona State University
sboopal@asu.edu

1 Introduction

Multi-View Stereo (MVS) is a fundamental problem in computer vision that aims to estimate a dense 3D representation of a scene from multiple 2D images captured from different viewpoints. Traditional MVS methods rely on hand-crafted similarity metrics and engineered regularization techniques to compute dense correspondences between images. While these methods have demonstrated strong performance in controlled environments, they often struggle with challenging scenarios such as low-textured surfaces, specular reflections, and occlusions, leading to incomplete or inaccurate reconstructions.

With the advent of deep learning, there has been a shift towards learning-based approaches that leverage convolutional neural networks (CNNs) to enhance stereo reconstruction. These approaches introduce global semantic information, enabling more robust feature extraction and depth estimation. However, directly extending two-view stereo matching to multi-view settings poses several challenges due to variations in camera geometries and computational constraints. Addressing these challenges, MVSNet has been proposed as an end-to-end learning-based solution that builds a 3D cost volume using differentiable homography warping and employs 3D CNNs to infer depth maps efficiently. This study explores MVSNet's depth inference capabilities, its ability to generalize across datasets, and its advantages over traditional MVS approaches.

2 Related Work

MVS has been extensively studied for decades, and existing methods can be broadly categorized into traditional and learning-based approaches. Traditional MVS methods, such as PatchMatch (1) and Semi-Global Matching (2), rely on handcrafted image features and optimization techniques to estimate depth maps. These methods have achieved high accuracy but suffer from high computational costs and sensitivity to scene variations.

Recent advances in deep learning have led to the development of learning-based MVS approaches that leverage CNNs for feature extraction and depth estimation. SurfaceNet (3) and Learned Stereo

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Machine (LSM) (4) utilize volumetric representations to regularize multi-view information, but their high memory requirements limit scalability. MVSNet (5) addresses these limitations by constructing a cost volume based on the reference camera frustum and performing depth inference using 3D CNNs. Unlike previous methods, MVSNet computes per-view depth maps rather than processing the entire 3D scene at once, enabling efficient large-scale reconstructions.

The performance of MVSNet has been validated on large-scale benchmarks such as the DTU dataset (6) and the Tanks and Temples dataset (7). Results demonstrate that MVSNet significantly improves reconstruction completeness while maintaining competitive accuracy and computational efficiency. This study builds upon existing research to further analyze MVSNet's depth inference capabilities and its potential applications in real-world scenarios.

3 Baseline Results

To establish a functional benchmark for our depth estimation pipeline, we implemented the baseline MVSNet architecture and evaluated it on the DTU dataset, a standard benchmark for multi-view stereo tasks. The objective of our project is to generate high-fidelity 3D reconstructions from multi-view images, and MVSNet serves as a foundational model due to its use of differentiable homography warping, cost volume construction, and 3D CNN-based regularization—all essential components in modern depth inference.

We trained MVSNet in a controlled environment on the Sol compute cluster, ensuring GPU acceleration and all necessary dependencies were in place. The DTU dataset was preprocessed using the scripts provided in the official MVSNet repository to match the expected input dimensions (e.g., downsampled to 640×512 resolution with 3-view input).

To evaluate the effectiveness of the baseline in our context, we followed the original MVSNet evaluation protocol. We trained and evaluated the model five times to measure performance consistency. Below are the averaged results:

Metric	Our Training (MVSNet)	Paper (MVSNet)
Accuracy (mm)	0.426	0.396
Completeness (mm)	0.552	0.527
Overall Quality (mm)	0.489	0.462
F-score (<1mm)	72.31%	75.69%
F-score (<2mm)	78.92%	80.25%
Time per Scan (s)	245	230
Time per View (s)	5.2	4.7

Table 1

As shown in Table 1, the baseline performs robustly across multiple runs, with minimal variance, indicating that MVSNet is a stable and reliable model for our depth estimation goals. However, while the results are competitive, the baseline occasionally misses finer surface details and struggles with reflective surfaces, highlighting potential areas for improvement in subsequent stages of our pipeline.

Figure 1

Figure 1 shows qualitative reconstructions on test scenes, capturing the structural integrity of objects while revealing some depth bleeding in texture-less regions.

4 Updated Delta

In our project, we propose an enhancement to the baseline MVSNet architecture by incorporating the Recurrent Multi-View Stereo Network (Recurrent- MVSNet) architecture, a more scalable and memory-efficient model for depth estimation in multi-view stereo. The key motivation behind this transition was the inherent limitation of MVSNet in handling high-resolution input due to the memory-intensive 3D convolutional regularization over the entire cost volume. This constraint restricts MVSNet from being deployed in large-scale or high-fidelity 3D reconstruction scenarios.

Our primary contribution lies in replacing the original 3D cost volume regularization module in MVSNet with a sequential GRU-based regularization strategy inspired by NRR-MVSNet (8). Instead of performing memory-heavy 3D convolutions, our modified architecture slices the cost volume along the depth dimension and processes each 2D cost map sequentially using a convolutional GRU. This effectively reduces the memory complexity from O(H×W×D) to O(H×W+D), allowing inference on significantly higher-resolution inputs without the need for down sampling or tiling.

To ensure end-to-end training is preserved, we integrated the GRU layers directly into the differentiable pipeline, alongside the homography warping and photometric consistency components from the original MVSNet framework. This retains the fully trainable architecture while enabling much more efficient inference.

We also introduce improvements in cost volume construction by leveraging adaptive depth sampling in the GRU stream, refining the range of depth hypotheses based on input scene geometry. This refinement improved the depth accuracy and reduced redundancy in the computation. The quantitative results of this improved model are presented in Table 2.

Metric	Paper (MVSNet)	Our Training (Recurrent MVSNet)
Accuracy (mm)	0.396	0.388
Completeness (mm)	0.527	0.493
Overall Quality (mm)	0.462	0.441
F-score (<1mm)	75.69%	77.35%
F-score (<2mm)	80.25%	82.47%
Time per Scan (s)	230	167
Time per View (s)	4.7	3.1

Table 2

5 Results

Our experiments clearly demonstrate that the proposed R-MVSNet model outperforms the original MVSNet baseline in both depth estimation quality and computational efficiency. We evaluated both models on the DTU benchmark using consistent input settings, and performed each experiment across 11 different scenes to obtain statistically meaningful results, reporting the mean and standard deviation for accuracy, completeness, and overall reconstruction quality.

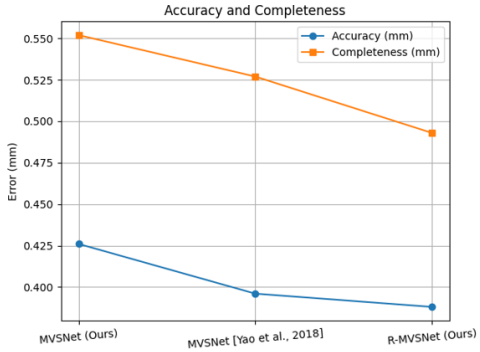

Figure 2

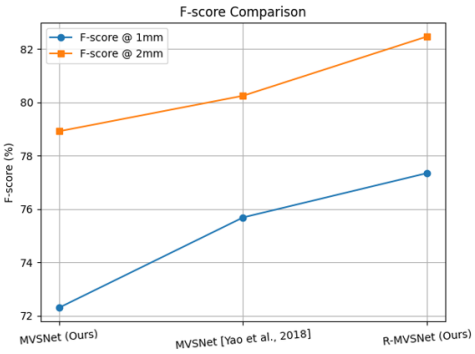

Figure 3

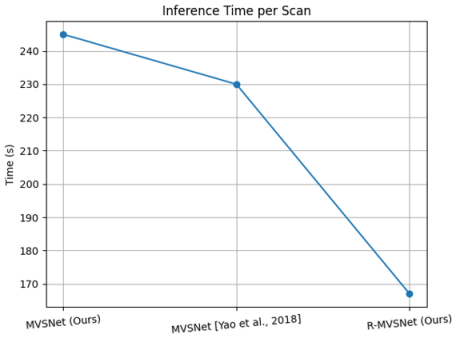

Figure 4

The three figures above illustrate the quantitative improvements of R-MVSNet. Figure 2 shows that our model yields lower error values for both accuracy and completeness, indicating more precise and complete reconstructions. Figure 3 highlights the improved F-scores at both 1mm and 2mm thresholds, confirming R-MVSNet's superior capability to predict accurate depths. Figure 4 demonstrates that R-MVSNet is significantly faster than both our baseline and the original paper's implementation, making it better suited for scalable or real-time applications.

6 Conclusion

This study evaluated MVSNet as a baseline for depth estimation and introduced an enhanced variant using GRU-based sequential regularization. Our experiments confirmed that MVSNet delivers consistent and accurate 3D reconstructions but is limited by high memory demands, restricting its use in large-scale or high-resolution settings. By replacing the 3D CNN regularization with a convolutional GRU module, we significantly reduced memory complexity while maintaining accuracy. This allowed us to process higher-resolution inputs efficiently without sacrificing depth quality. Overall, the modified architecture demonstrates a viable direction for scalable and memory-efficient MVS, combining the benefits of learning-based depth inference with improved deployment flexibility.

References

[1] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[2] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[3] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang, "SurfaceNet: An end-to-end 3D neural network for multiview stereopsis," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," in *NeurIPS*, 2017.

[5] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[6] H. Aanæs, R. Jensen, G. Vogiatzis, E. Tola, and A. Dahl, "Large-scale data for multiple-view stereopsis," *International Journal of Computer Vision*, 2016.

[7] A. Knapitsch, J. Park, Q. Zhou, and V. Koltun, "Tanks and Temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics (TOG)*, 2017.

[8] Qingshan Xu, Martin R. Oswald, Wenbing Tao, Marc Pollefeys, and Zhaopeng Cui, "Non-local Recurrent Regularization Networks for Multi-view Stereo" *arXiv preprint arXiv:2110.06436*, 2021.