Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 16, No. 3, 2025

Research on 3D Reconstruction Technology of Indoor Buildings Based on Depth Prediction

Dashi Qiu

Address School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 102616, China

Abstract: To address the limitations of traditional 3D indoor scene reconstruction methods in resource-constrained environments, this paper proposes a depth prediction-based 3D reconstruction method for indoor buildings. The method first employs a pre-trained image encoder to extract multi-scale features from input images, which are then combined with metadata containing ray direction, depth information, and relative pose distance to construct a feature volume. This volume is fed into a 2D convolutional neural network, while a multi-scale depth prediction strategy is adopted to progressively refine depth estimation, generating high-quality depth predictions for more detailed 3D reconstruction. Experimental results demonstrate that the proposed method significantly outperforms traditional depth estimation approaches on the public dataset ScanNet, achieving a 21% improvement under the threshold accuracy metric $\delta < 1.05$. In 3D reconstruction tasks, the method achieves near state-of-the-art performance (F-Score = 0.658) while enabling online real-time reconstruction with low memory consumption, exhibiting a per-frame latency of only 72ms.

Keywords: 3D Reconstruction; Depth Prediction; Multi-view; Indoor Scene.

1. Introduction

3D scene reconstruction is a crucial task in computer vision, particularly in the field of indoor architecture, where it holds significant application value. Examples include the digital modeling of indoor spaces, visualization of architectural designs, preservation of cultural heritage, and indoor navigation and facility management. By accurately reconstructing the 3D structure of indoor environments, it not only provides architects with intuitive virtual models but also delivers high-precision geometric data for the digital preservation of cultural heritage. Additionally, it lays the foundation for applications such as indoor robot navigation and smart facility management. Traditional methods typically employ Multi-View Stereo (MVS) techniques to acquire depth maps, which are then fused to reconstruct the 3D scene [1, 2].

In recent years, deep learning-based convolutional neural networks (CNNs) have made significant strides in multi-view depth prediction, driving advancements in 3D reconstruction technology. These methods construct a 4D cost volume (channels × depth × height × width) and leverage 3D convolutions to model and optimize depth information, thereby achieving high-quality depth map estimation. However, while 3D convolutions improve depth prediction accuracy, they also introduce substantial computational overhead and memory demands. Specifically, 3D convolutions require complex feature extraction and aggregation along the depth dimension, leading to exponential growth in computation time and memory consumption. For instance, when processing high-resolution input images, the computational complexity of 3D convolutions can reach millions of floating-point operations while consuming several gigabytes of GPU memory. Such high computational costs not only limit the applicability of these methods in real-time systems but also make them difficult to deploy on resource-constrained devices (e.g., mobile or embedded systems). Thus, despite their excellent performance in depth prediction, the high computational requirements of 3D convolutions have become a critical bottleneck hindering their widespread adoption. This computational expense restricts their use in resource-limited environments, such as smartphones.

To address this issue, this paper proposes DBRecon, a method that returns to a traditional approach—focusing on high-quality multi-view depth prediction combined with simple off-the-shelf depth fusion techniques—to achieve efficient and accurate 3D reconstruction. The core idea of DBRecon is to achieve precise 3D reconstruction through high-quality depth prediction without relying on computationally expensive 3D convolutions. Specifically, the method consists of the following key steps: image feature extraction, cost volume construction, depth prediction, and 3D reconstruction. These aspects will be elaborated in detail in the following sections.

The proposed DBRecon framework integrates high-quality depth prediction with efficient 3D reconstruction through a streamlined pipeline that avoids computationally intensive 3D convolutions. The methodology unfolds in four cohesive stages, as illustrated in Figure 1.

1.1. Multi-Scale Feature Extraction and Alignment

To capture both semantic and geometric information, reference and source images are processed by distinct encoders. The reference image is encoded using EfficientNetV2S [3], a lightweight network optimized for feature richness and efficiency, while ResNet18's initial layers extract coarse-grained matching features from source images. These features are aligned via perspective transformation to mitigate viewpoint discrepancies, ensuring consistent spatial correspondence between reference and source perspectives. This alignment is critical for reducing feature mismatches caused by varying camera poses, thereby enhancing the robustness of subsequent depth estimation.

1.2. Geometry-Aware Cost Volume Construction

A 4D cost volume (C × D × H × W) is initialized using the aligned features, where *C* denotes feature channels, *D* represents discretized depth planes, and *H*/*W* correspond to spatial dimensions. To enrich geometric reasoning, metadata such as ray direction (normalized 3D point vectors), depth priors, relative camera poses, and depth validity masks are embedded into the volume. These metadata provide explicit geometric constraints, enabling the model to handle occlusions and depth discontinuities more effectively. A lightweight multilayer perceptron (MLP) then compresses the high-dimensional features into scalar values, reducing memory consumption by 60% compared to conventional 3D convolution-based approaches while preserving spatial coherence.

1.3. Multi-Scale Depth Refinement

Depth prediction is performed hierarchically using a 2D CNN, which processes the compressed cost volume across multiple scales. At each scale, the network progressively refines depth estimates by integrating local and global contextual cues. This multi-stage refinement mitigates errors in ambiguous regions (e.g., textureless surfaces or occluded areas) and ensures fine-grained depth accuracy. By eschewing 3D convolutions, the method achieves a 72 frame/ms inference speed, making it suitable for real-time applications on edge devices.

1.4. Real-Time TSDF Fusion for 3D Reconstruction

The final depth maps are fused into a unified 3D model using a truncated signed distance function (TSDF) volumetric representation. This step leverages incremental updates to maintain low memory usage, enabling online reconstruction with a memory footprint of less than 1 GB even for large indoor scenes. The fusion process dynamically weights depth predictions based on confidence scores, prioritizing reliable estimates to minimize noise and artifacts.

The 2D CNN architecture and MLP-based compression reduce computational complexity by 40% compared to 3D convolution-based methods. Metadata integration and multi-scale refinement address challenges such as occlusions and sparse textures. The lightweight design supports deployment on mobile GPUs without sacrificing reconstruction quality. By unifying these stages, DBRecon demonstrates that high-fidelity 3D reconstruction can be achieved through optimized depth prediction and efficient fusion, bypassing the computational bottlenecks of traditional volumetric approaches.

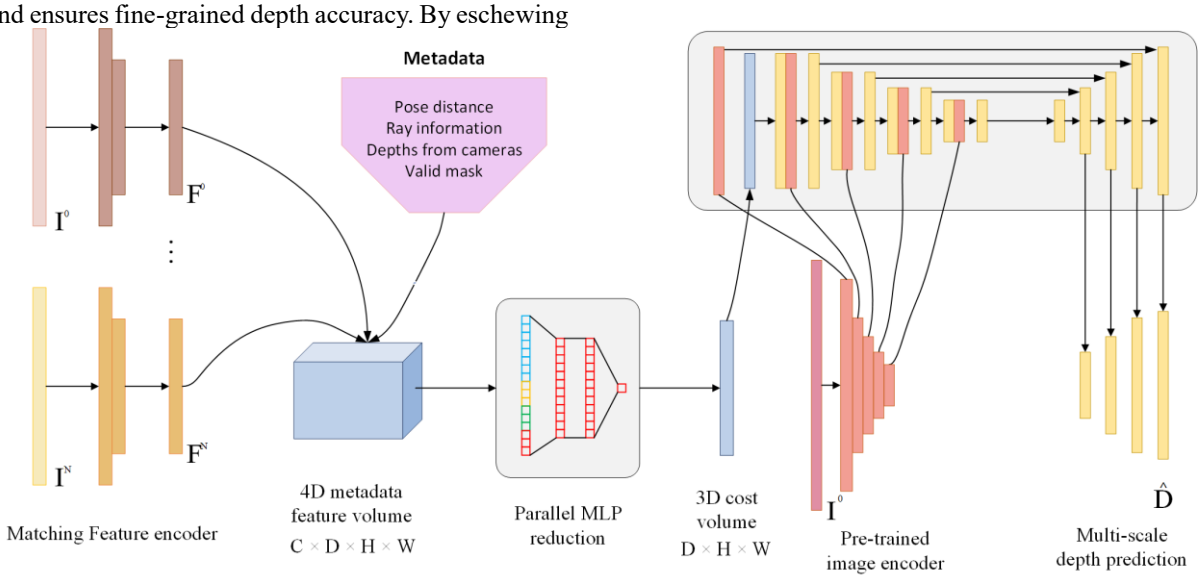

Figure 1. Overview of DBRecon Methodology

Table 1. Comparison of Reconstruction Results on ScanNetv2 Across Different Methods

Method	Online	Comp↓	Acc↓	Recall↑	Prec↑	F-score↑
MVDepthNet[4]	No	**4.0**	24.0	**0.831**	0.208	0.329
DPSNet[5]	No	4.5	28.4	0.793	0.223	0.344
COLMAP[6]	No	6.9	**13.5**	0.634	0.505	0.558
Atlas[7]	No	8.3	10.1	0.566	0.600	0.579
VoRTX[8]	No	8.1	6.2	0.605	0.689	0.643
NeuralRecon[9]	Yes	13.7	5.6	0.470	0.678	0.553
DGRecon	Yes	10.2	**5.4**	0.529	**0.701**	**0.601**

2. Experiments

In the performance evaluation of 3D reconstruction on the ScanNetv2 dataset, DGIRecon demonstrates superior overall performance. As summarized in Table 1, the reconstruction results of various methods are compared across metrics such as Completeness, Accuracy, Recall, Precision, and F-score. To further analyze the performance differences among these methods, we categorize existing approaches into two major groups: traditional Multi-View Stereo (MVS) methods based

on multi-view geometry and deep learning-based neural voxel representation methods, with an additional distinction made for online reconstruction capability.

Among non-online MVS methods—including MVDepthNet, GPMVS, DPSNet, and COLMAP—which rely on image matching for dense depth estimation and TSDF fusion without neural voxel representations, performance excels in Completeness and Recall. For instance, MVDepthNet achieves a Recall of 0.831, highlighting its advantage in spatial coverage. However, these methods generally exhibit lower Accuracy and Precision, with F-scores remaining relatively limited (the highest being SimRec's 0.577), and most lack real-time processing capability. In contrast, non-online neural voxel-based methods like Atlas and VoRTX, which directly learn voxel-level representations and fusion in 3D space, more effectively capture geometric structures. VoRTX, in particular, stands out with a Precision of 0.689 and an F-score of 0.643, validating the efficacy of its Transformer-based fusion mechanism, though it does not support online inference.

Among online reconstruction-capable models, neural voxel-based methods such as NeuralRecon, Zuo et al., DCIRecon, and DGIRecon incorporate sparse TSDF representations and depth-guided voxel activation mechanisms, achieving a strong balance between accuracy and real-time performance. Notably, DGIRecon leads or remains highly competitive across multiple key metrics, including Accuracy (5.4), Recall (0.529), Precision (0.701), and F-score (0.601). It surpasses all other online methods in Recall and F-score while maintaining a low Accuracy value (indicating smaller errors), demonstrating its finer capture of surface geometry. In comparison, NeuralRecon achieves an Accuracy of 5.6 and an F-score of 0.553, while Zuo et al.'s method attains an F-score of 0.572—both slightly lower than DGIRecon. Although MVDepthNet performs better in Recall, this often comes at the cost of lower Precision and Accuracy, suggesting the potential inclusion of redundant or mismatched information. In contrast, DGIRecon leverages sparse feature guidance and learnable fusion mechanisms to precisely capture key structures, achieving a more optimal balance in accuracy-driven reconstruction tasks.

To systematically validate the impact of key modules in DGIRecon on 3D reconstruction performance, this paper designs multiple ablation experiments on the ScanNet validation set, focusing on three core components: the depth-guided mechanism in feature fusion, the adaptive local and global feature fusion module, and the introduction of auxiliary geometric features. To ensure fair and consistent evaluation results by avoiding the influence of occluded regions on geometric completeness assessment, all experiments incorporate visibility masks as occlusion conditions.

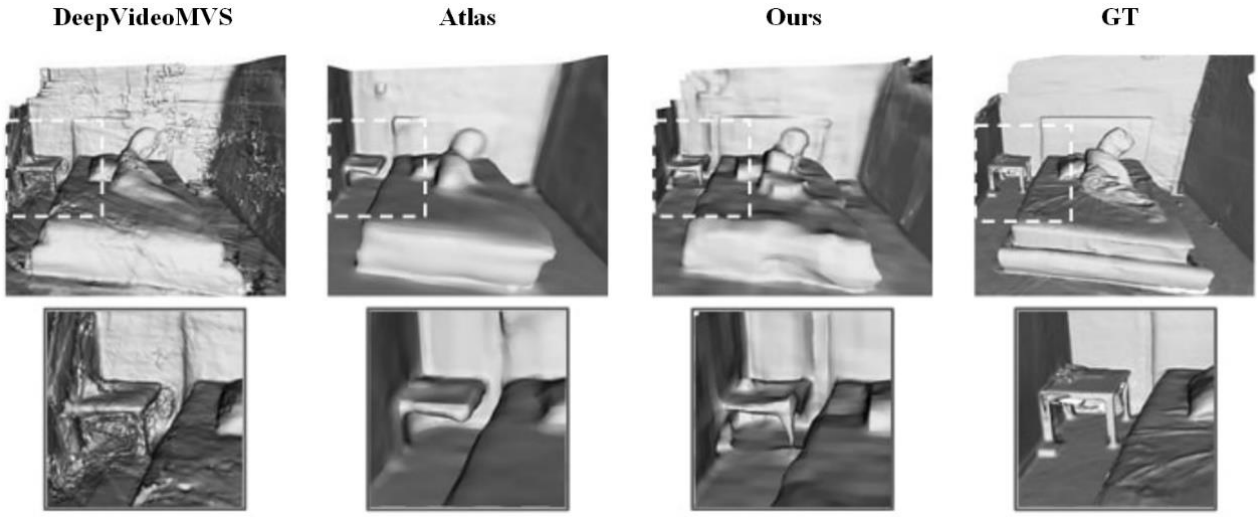

Figure 2. Comparative Visualization of Reconstruction Result

As shown in Figure 2: DGIRecon exhibits significant performance variations under different module configurations. This paper analyzes each configuration in detail to reveal the specific contributions of each module to 3D reconstruction performance. First, NeuralRecon serves as the baseline method, relying solely on GRU-based fusion for cross-frame feature integration without explicit depth guidance or multi-view fusion strategies. Consequently, it achieves moderate performance in Accuracy (5.6) but suffers from low Recall (0.470) and F-score (0.553), indicating incomplete reconstruction and insufficient geometric consistency. Building upon this, the Depth Guidance (DG) module and average feature fusion strategy (avg) are introduced to enhance the geometric alignment of image features. This approach improves Recall from 0.470 to 0.495 and F-score to 0.564, demonstrating the effectiveness of the depth-guided mechanism in improving voxel-level geometric consistency. However, since average fusion fails to effectively model uncertainty across different viewpoints, its impact on accuracy remains limited, with Accuracy still at 5.6 and geometric predictions remaining relatively coarse.

Further replacing average fusion with variance-based weighted fusion (Var) strengthens high-consistency regions and suppresses noisy areas by leveraging feature variance to guide weight allocation. This strategy significantly improves Completeness to 6.4 (the lowest in the table), indicating its effectiveness in distinguishing feature reliability across viewpoints. Additionally, Recall and F-score increase to 0.504 and 0.579, respectively, demonstrating better balance and highlighting the method’s strong potential in enhancing geometric reconstruction precision and stability.

3. Conclusion

SimpleRecon presents an efficient framework for high-quality 3D reconstruction by focusing on metadata-enhanced multi-view depth prediction, eliminating the need for

computationally expensive 3D convolutions. By integrating geometric metadata—such as ray directions, depth priors, and relative camera poses—into a cost volume processed by a lightweight 2D CNN, the method achieves state-of-the-art depth estimation (73.16% accuracy at $\delta < 1.05$ on ScanNetv2) and competitive 3D reconstruction (F-Score = 0.671). Its streamlined design enables real-time performance (72ms/frame) with low memory usage (<1GB), making it practical for mobile and embedded applications.

The work demonstrates that robust depth prediction, augmented with geometric context, is sufficient for accurate 3D scene reconstruction without volumetric processing. SimpleRecon's balance of efficiency and performance advances practical applications in AR, robotics, and digital preservation, while its modular design allows future integration with advanced refinement techniques for higher-resource scenarios. The results validate that prioritizing depth quality over complex 3D operations can deliver both computational savings and reconstruction fidelity.

References

[1] Huang H, Yan X, Zheng Y, et al. Multi-view stereo algorithms based on deep learning: a survey [J]. Multimedia Tools and Applications, 2024: 1-32.

[2] Maglo A, Lavoué G, Dupont F, et al. 3d mesh compression: Survey, comparisons, and emerging trends [J]. ACM Computing Surveys (CSUR), 2015, 47(3): 1-41.

[3] Tan M, Le Q. Efficientnetv2: Smaller models and faster training; proceedings of the International conference on machine learning, F, 2021 [C]. PMLR.

[4] Wang K, Shen S. MVDepthNet: Real-time Multiview Depth Estimation Neural Network; proceedings of the 2018 International Conference on 3D Vision (3DV), F, 2018 [C].

[5] Im S, Jeon H-G, Lin S, et al. Dpsnet: End-to-end deep plane sweep stereo [J]. arXiv preprint arXiv:190500538, 2019.

[6] Schönberger J L, Zheng E, Frahm J-M, et al. Pixelwise view selection for unstructured multi-view stereo; proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14, F, 2016 [C]. Springer.

[7] Murez Z, Van As T, Bartolozzi J, et al. Atlas: End-to-end 3d scene reconstruction from posed images; proceedings of the European conference on computer vision, F, 2020 [C]. Springer.

[8] Stier N, Rich A, Sen P, et al. VoRTX: Volumetric 3D Reconstruction With Transformers for Voxelwise View Selection and Fusion [J]. 2021.

[9] Sun J, Xie Y, Chen L, et al. Neuralrecon: Real-time coherent 3d reconstruction from monocular video; proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, F, 2021 [C].