실내 정밀 3D 재구성을 위한 깊이 인식 Transformer 기반 TSDF 예측 기법

문명운*, 조경은**
* 동국대학교 NUI/NUX 플랫폼 연구센터 연구원
** 동국대학교 첨단융합대학 컴퓨터 AI 학부 교수

wmy_dongguk@dongguk.edu, cke@dongguk.edu(교신저자)

Depth-aware transformer-enhanced TSDF prediction for Detailed Indoor 3D Reconstruction

Mingyun Wen*, Kyungeun Cho**
* NUI/NUX Platform Research Center, Dongguk University
** Dept. of Computer Science and Artificial Intelligence, College of Advanced Convergence Engineering, Dongguk University

Abstract

3D scene reconstruction from monocular videos plays an important role in robotics navigation and VR/AR applications. Recent advances in deep learning have significantly improved reconstruction quality. However, a performance gap remains compared to methods that leverage direct depth data, such as LiDAR point clouds, particularly in capturing fine geometric details. Monocular video-based approaches often produce over-smooth surfaces and fail to recover thin structures. To address this, we proposed a depth-aware Transformer-based module that aggregates multi-view image features and assigns adaptive importance to each view via attention mechanism. Integrated into the TSDF prediction stage of Finerecon, our method enhances geometric fidelity by leveraging both multi-view features and estimated depth information. Experiments on ScanNet v2 dataset demonstrate that our approach successfully recovers fine surface details and improves reconstruction quality compared to the baseline.

1. Introduction

Reconstructing detailed indoor geometry from monocular videos is fundamental to AR/VR, robotics, and digital twins [1-2]. Traditional handcrafted feature-based approaches typically produce sparse point clouds and struggle with complex scenes. Thanks to the strong representation capability of deep learning, dense 3D geometry estimation from images has become feasible. Nevertheless, monocular video-based methods still exhibit a significant performance gap compared to approaches that use direct 3D data as input, such as depth maps or point clouds.

FineRecon [1] leverages a depth estimation network to obtain per-view depth maps, combines them with image features, and predicts voxel occupancy and TSDF values. This design improves global consistency and completeness of reconstructed geometry. However, its uniform fusion of multi-view features often leads to over-smoothed surfaces, since occlusion, lighting, and visibility cause different views to contribute unequally to surface details.

To overcome this limitation, we propose a lightweight attention module for multi-view feature integration. Our method introduces a depth-aware Transformer into the TSDF prediction stage of FineRecon. By predicting per-view attention weights and incorporating depth information, the module adaptively emphasizes informative views and suppresses noisy ones. As a result, the reconstructed meshes better preserve local geometric details such as thin structures and fine textures.

2. Proposed method

The overview of the network architecture is as shown in Figure 1. Given input RGB keyframes, we extract both coarse and fine-scale 2D features using two conventional branches. The coarse branch produces feature maps $\{\mathcal{F}^c\}$ that preserve global context, the fine branch generates $\{\mathcal{F}^f\}$ that focus on local geometric details. Additionally, a depth map for per keyframe is estimated using pretrained depth estimation networks [2]. These depth maps are used to form a fused TSDF

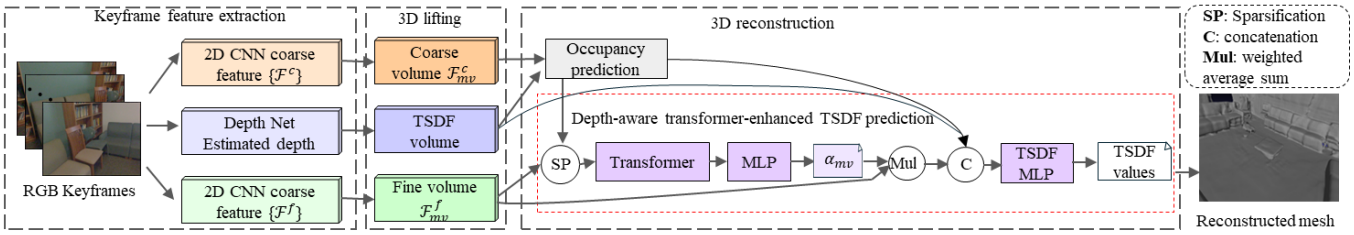

Figure 1 overview of architecture of proposed method

volume, following the same integration strategy as Finerecon.

Each 2D feature map is projected into the 3D voxel space using known camera poses and depth-based lifting. This produces coarse feature volume $\{\mathcal{F}^c_{mv}\}$, fine feature volume $\{\mathcal{F}^f_{mv}\}$, and TSDF volume from estimated depths. These volumes are voxel-aligned and serve as inputs to the subsequent geometry prediction modules.

The coarse feature volume $\{\mathcal{F}^c_{mv}\}$, is used to predict voxel occupancy probabilities. This branch follows the baseline FineRecon design and provides a global constraint for geometry.

To refine surface details, we propose a Transformer-based multi-view fusion module that incorporates depth-aware cues. First, the fine feature volume $\{\mathcal{F}^f_{mv}\}$ is sparsified (SP) and processed by a Transformer, followed by an MLP, to predict per-view attention weights α_{mv}. These weights are applied via element-wise multiplication to the fine features, followed by weighted averaging across views. The aggregated feature is concatenated (C) with the fused TSDF volume, producing a joint representation. A TSDF regression MLP then predicts the final voxel-wise signed distance values. This design explicitly leverages both multi-view image features and depth-derived TSDF priors, enabling more accurate surface localization compared to purely image-feature fusion. The predicted TSDF values are converted into a triangular mesh via marching cubes [3].

3. Experimental results

We conducted experiments on the ScanNet v2 [4] dataset to evaluate our proposed method. Following the same loss functions and training settings as FineRecon, the model was trained on an Ubuntu server equipped with six NVIDIA RTX A6000 GPUs.

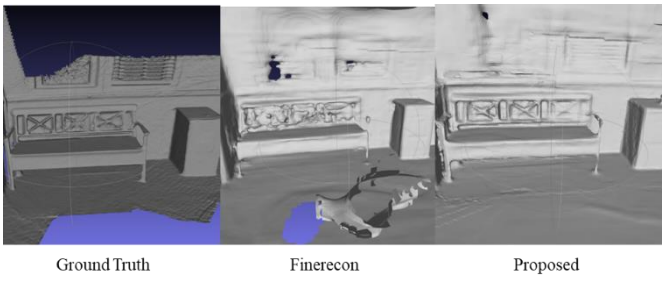

Figure 2 Experimental results on ScanNet v2

As shown in Figure 2, compared to FineRecon, our approach recovers finer geometric structures and produces sharper object boundaries. In particular, the backrest and thin structures of the sofa are better preserved, and the surfaces appear less over-smoothed. These results demonstrate that the proposed depth-aware Transformer module effectively enhances local detail reconstruction.

4. Conclusion

We presented a depth-aware Transformer-enhanced TSDF prediction method for indoor 3D reconstruction. Experiments on ScanNet v2 show that our approach improves the recovery of fine geometric details compared to the baseline.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2022-NR070225) (90%) and Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2025-RS-2023-00254592) grant funded by the Korea government (MSIT) (10%).

References

[1] N. Stier et al., "Finerecon: Depth-aware feed-forward network for detailed 3d reconstruction," In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, Oct. 2-6 2023, pp. 18423-18432

[2] M. Sayed, J. Gibson, J. Watson, V. Prisacariu, M. Firman, and C. Godard, "Simplerecon: 3d reconstruction without 3d convolutions," In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, Oct. 23-27 2022, pp. 1-19

[3] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," in Seminal graphics: pioneering efforts that shaped the field, 1998, pp. 347-353.

[4] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, Jul. 21-26 2017, pp. 5828-5839.