Real-Time Temporally Consistent Depth Completion for VR-Teleoperated Robots

Chengfan Li[1] Automne Petitjean[2] Are Oelsner[1] Stefanie Tellex[1] James Tompkin[1]

[1]Brown University [2]ENS de Lyon

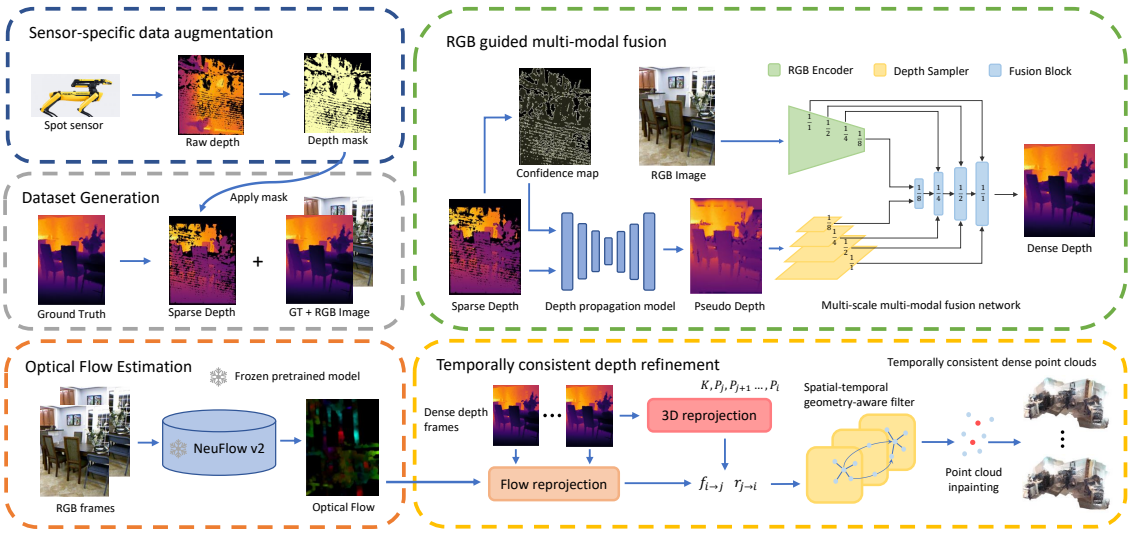

Figure 1. Overview of the proposed method.

1. Introduction

High-quality visual perception is essential for precise interaction in VR-teleoperated robots. Existing systems are challenged by sparse inputs and high latency, emphasizing the need for real-time, temporally consistent, and dense point cloud reconstruction. In this paper, we present a real-time depth completion and point cloud reconstruction system for VR-teleoperated robots. We employ an algebraically-constrained, normalized CNN to propagate depth and confidence through multi-scale multi-modal fusion network regulated by a gradient matching loss [3]. We also implement a spatial-temporal geometry-aware filter to ensure temporally consistent point cloud reconstruction. This system achieves a rendering speed of 50 FPS and demo videos are available at project page.

1.1. Sensor-Specific Data Augmentation

We employ sensor-specific data augmentation that aligns with the spot's sensor input distribution by applying a spot-collected sensor mask to generate training input for the model. Our experiments demonstrate that this sensor-specific augmentation significantly improves depth completion accuracy and visual quality on the spot's sensor.

1.2. Real-Time Depth Completion

Unguided Propagation Model Building on the architecture presented in Nconv [1], we introduce an propagation model that integrates algebraically-constrained normalized convolution layers to propagate depth and confidence to subsequent layers. By generating pseudo depth maps prior to multi-modal fusion, the propagation model addresses the challenges posed by sparse data, improving fusion quality and accelerating the convergence of the fusion network.

RGB Guided Multi-Scale Multi-Modal Fusion After propagating sparse data to generate pseudo depth maps, we incorporate RGB data from the spot's camera to enhance edge, corner and texture information. Initially, an encoder extracts features from the RGB images. These RGB features are then concatenated with downsampled depth maps using an early fusion strategy and fed into the multi-modal fusion decoder. We employ a multi-scale approach, integrating information from coarse to fine levels. This fusion process results in the final high-resolution depth map, improving depth completion accuracy.

1.3. Temporally Consistent Depth Refinement

The robot and its surrounding environment are dynamic, leading to noisy and unstable point clouds. To address this, after generating dense depth maps, we apply a spatial-temporal geometry-aware filter to produce consistent point clouds and smooth transitions. This approach optimizes the rendering and reconstruction quality of dynamic scenes.

Optical Flow and Pose Estimation NeuFlow-v2 [4] is employed for online optical flow estimation between frames. To complement the 2D motion information from optical flow, the built-in kinematics module in ROS is used to obtain real-time pose estimations.

Spatial-Temporal Geometry-Aware Filter A depth filter based on flow trajectories proposed by RobustCVD [2] resolves fine-scale depth details. We extend this method to 3D space by leveraging the relative pose between frames, integrating original flow-based reprojection, and applying weighted spatial-temporal smoothing to produce temporally consistent point clouds and seamless transitions.

References

[1] Abdelrahman Eldesokey, Michael Felsberg, and Fahad Shahbaz Khan. Confidence propagation through cnns for guided sparse depth regression. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2423–2436, 2019. 1

[2] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021. 1

[3] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 1

[4] Zhiyong Zhang, Aniket Gupta, Huaizu Jiang, and Hanumant Singh. Neuflow v2: High-efficiency optical flow estimation on edge devices. *arXiv preprint arXiv:2408.10161*, 2024. 1