RIAV-MVS: Recurrent-Indexing an Asymmetric Volume for Multi-View Stereo

Changjiang Cai, Pan Ji, Qingan Yan, Yi Xu
OPPO US Research Center, InnoPeak Technology, Inc.

Scan for Code!

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Overview

- **RIAV-MVS** as a "learning-to-optimize" method for multi-view depth estimation from posed images:
  - A new paradigm to predict the depth via learning to recurrently index an asymmetric plane-sweeping cost volume via GRUs
  - A residual pose module to correct the relative poses between images/cameras

Motivation

- **Existing CNN-based MVS methods**:
  - Siamese CNN-based encoder for feature learning:
    - 1. Symmetric features for the reference and the source images
    - 2. CNN-based encoder being short of global context
  - Plane-sweeping cost volume via differentiable homography:
    - 3. Assuming pose being accurate
  - 3D-CNN encoder-decoder for cost volume regularization:
    - 4. Time and memory consuming by 3D-CNNs
    - 5. Soft-Argmin for depth regression not robust to multi-modal distributions

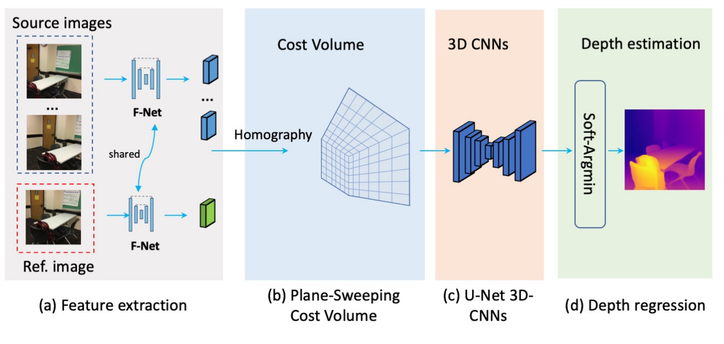


Fig. 1

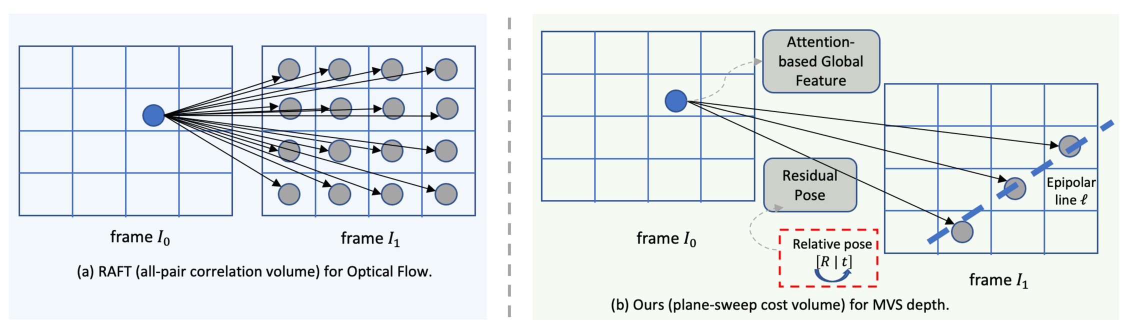


Fig. 2

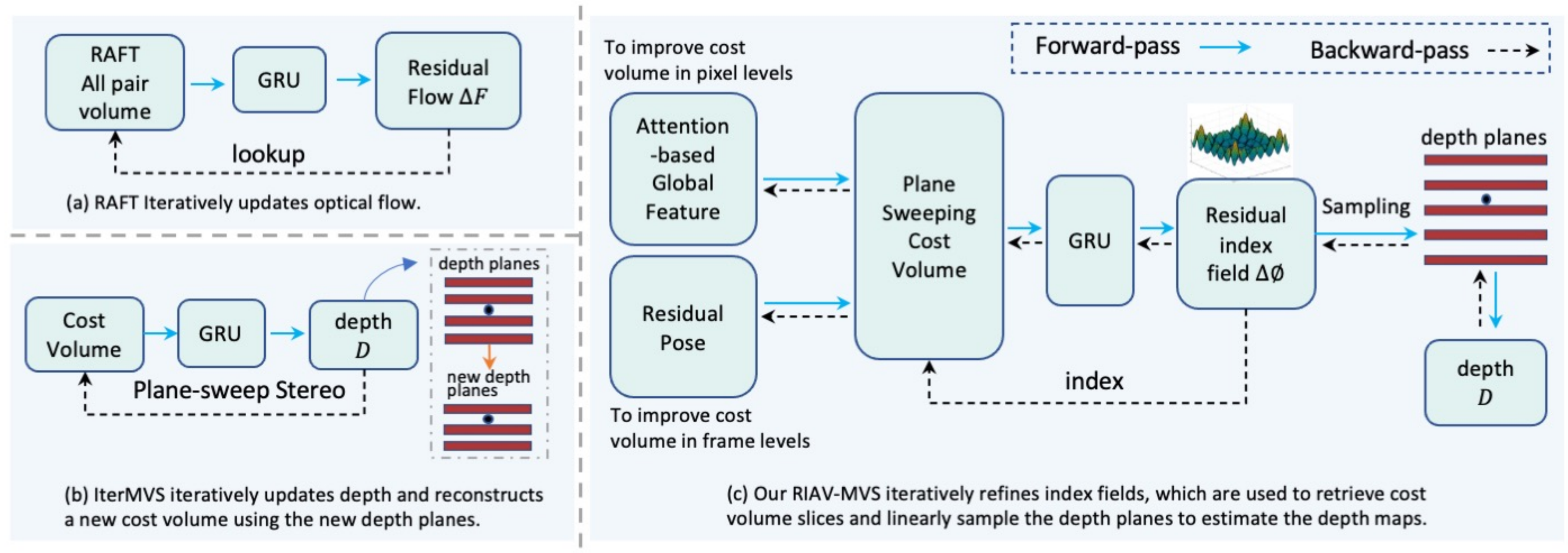


Fig. 3

Background

- Our RIAV-MVS vs RAFT(Teed & Deng):
  - Borrowing ideas from RAFT(Teed & Deng) for learning to optimize via GRU that performs lookups on the correlation volumes with non-trivial modifications (Fig. 2):
    - RAFT's all-pair correlation for optical flow: no multi-view geometry constraints → plane-sweeping cost volume for MVS
    - Our proposed ***index filed*** serves as a new design to bridge cost volume optimization and depth map estimation
- Our RIAV-MVS vs IterMVS (Wang et. al):
  - IterMVS predicts the depth and reconstructs a new plane-sweep cost volume using updated depth planes (Fig. 3.b)
  - Ours learns to index the cost volume by approaching the "correct" depth planes per pixel via an index field (Fig. 3.c).

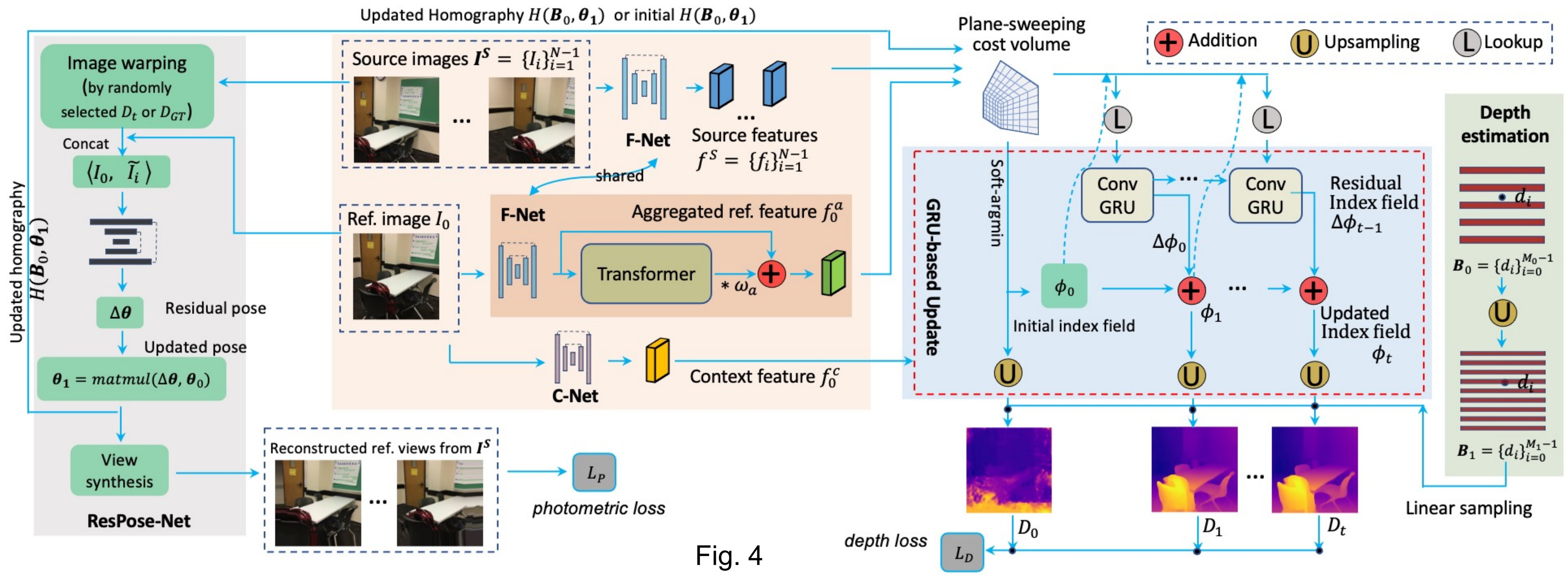


Fig. 4

Approach

- Our proposed network consists of
  - Feature extraction (i.e., F-Net, a Transformer, and C-Net) blocks
  - Cost volume construction
  - Index field GRU-based optimization and
  - Residual pose update
- We propose to improve the cost volume at pixel- and frame- levels

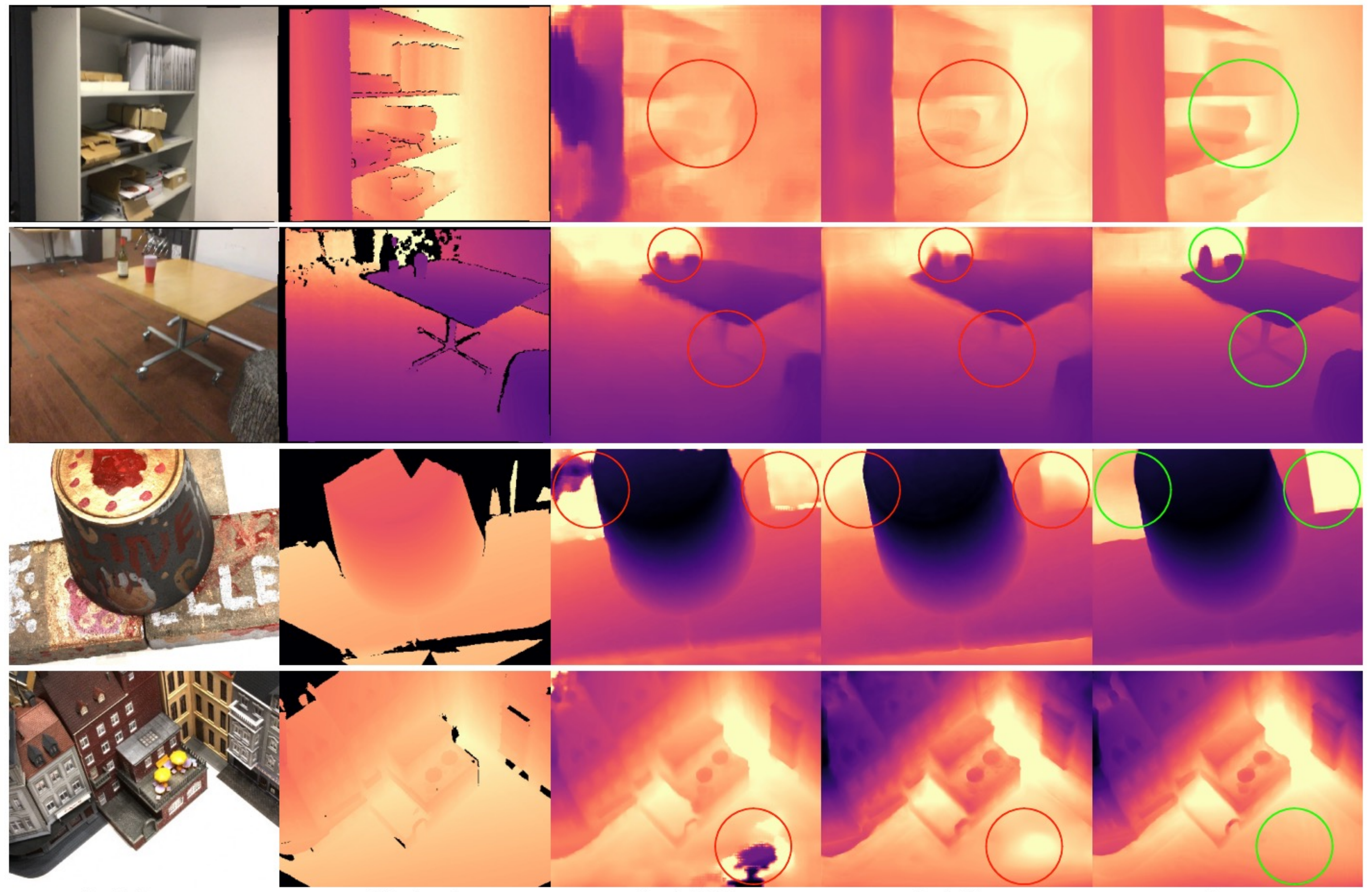
Approach (Cont.)

- At the pixel level, a transformer block is asymmetrically applied to the reference view (but not to the source views): Using global context via a transformer and pixel-wise local CNN features, an ***asymmetric*** cost volume is constructed
- At the frame level, a residual pose net to rectify the camera poses: the rectified poses are used to more accurately warp the reference features to match the counterparts in source views

Results

- Depth map results on ScanNet and DTU

Model Variants	ScanNet Test-Set		
	Abs-Rel	Abs (meters)	δ < 1.25
Our (base)	0.0885	0.1605	0.9211
Our (+pose)	0.0827	0.1523	0.9277
Our (+pose,atten)	**0.0734**	**0.1381**	**0.9395**



Ref. image GR depth IterMVS PairNet Ours