NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video

Jiaming Sun*  Yiming Xie*  Linghao Chen  Xiaowei Zhou  Hujun Bao
CVPR 2021 (Oral and Best Paper Candidate)
* equal contribution
Motivation

3D Geometry is crucial for immersive AR effects

3D Reconstruction

Credit: DepthLab, Apple Clips with LiDAR
Motivation

Real-time 3D Geometry is crucial for immersive AR effects

3D Reconstruction

Credit: DepthLab, Apple Clips with LiDAR
Motivation

Depth sensor v.s. Monocular camera

With a depth sensor

😊 Accurate depth measurement
😊 Takes a lot of energy
😊 Only available to a few high-end products

Credit: RealSense, Azure Kinect, LiDAR on iPad Pro
Motivation
Depth sensor v.s. Monocular camera

**With a depth sensor**

- 😃 Accurate depth measurement
- 😞 Takes a lot of energy
- 😞 Only available to a few high-end products

**With a monocular camera**

Credit: RealSense, Azure Kinect, LiDAR on iPad Pro
Motivation

Depth sensor v.s. Monocular camera

With a depth sensor

😄 Accurate depth measurement
😊 Takes a lot of energy
🙂 Only available to a few high-end products

With a monocular camera

😊 Immediately available to many phones
😊 Not as accurate as depth sensors
😊 Not as fast as depth sensors

Credit: RealSense, Azure Kinect, LiDAR on iPad Pro
Motivation

Depth sensor v.s. Monocular camera

With a depth sensor

😊 Accurate depth measurement
😊 Takes a lot of energy
😊 Only available to a few high-end products

With a monocular camera

😊 Immediately available to many phones
😊 Not as accurate as depth sensors
😊 Not as fast as depth sensors

Credit: RealSense, Azure Kinect, LiDAR on iPad Pro
Motivation
Pipeline overview for depth-based methods
Motivation

Pipeline overview for depth-based methods

Key-frame Selection → Multi-view Depth Estimation → TSDF Fusion

Depth Sensor + KinectFusion

Newcombe et al., "KinectFusion: Real-Time Dense Surface Mapping and Tracking", UIST 2011
Motivation

MVSNet: Depth Inference for Unstructured Multi-view Stereo

Yao Yao¹, Zixin Luo¹, Shiwei Li¹, Tian Fang², and Long Quan¹
Motivation

DeepTAM: Deep Tracking and Mapping
Huizhong Zhou* Benjamin Ummenhofer* Thomas Brox

DEEPV2D: VIDEO TO DEPTH WITH DIFFERENTIABLE STRUCTURE FROM MOTION
Zachary Teed
Princeton University
zteed@cs.princeton.edu
Jia Deng
Princeton University
jiadeng@cs.princeton.edu

BA-NET: DENSE BUNDLE ADJUSTMENT NETWORKS
Chengzhou Tang
School of Computer Science
chengzhou_tang@sfu.ca
Ping Tan
School of Computer Science
pingtan@sfu.ca

MVD depthNet: Real-time Multiview Depth Estimation Neural Network
Kaixuan Wang Shaojie Shen
Hong Kong University of Science and Technology

Neural RGB→D Sensing: Depth and Uncertainty from a Video Camera
Chao Liu1,2* Jinwei Gu1,3* Kihwan Kim1 Srinivasa Narasimhan2 Jan Kautz1
1NVIDIA 2Carnegie Mellon University 3SenseTime

Recently: Cascade-Stereo, DeepSFM, CNMNet, Consistent Depth...
Motivation

Depth-based methods v.s. NeuralRecon

Depth-based methods

 EITHER LAYERED OR SCATTERED RESULTS
Motivation

Depth-based methods v.s. NeuralRecon

Depth-based methods

😊 Either layered or scattered results
😊 Redundant computation
Motivation

Depth-based methods v.s. NeuralRecon

Depth-based methods

 EITHER LAYERED OR SCattered RESULTS
 REDUNDANT COMPUTATION

Our solution: NeuralRecon

 RECONSTRUCT LOCAL SURFACES DIRECTLY IN TSDF
 JOINT FRAGMENT RECONSTRUCTION AND FUSION
 BETTER QUALITY AND FASTER SPEED
NeuralRecon

Overview

Input: Posed Images

End-to-End System

Output: Scene Geometry (sparse TSDF volume)
NeuralRecon

Overview

Input: Posed Images

→ Key-frame Selection

→ Fragment Reconstruction

→ Fuse to Global Volume

Output: Scene Geometry (sparse TSDF volume)
NeuralRecon
Fragment reconstruction

Input: Posed Images in a fragment

Feature Unprojection

View-Independent Volume

Sparse Conv

Fragment Geometry (sparse TSDF volume)

Fragment Reconstruction

...
NeuralRecon
Fragment reconstruction

Input:
Posed Images in a fragment

Feature Unprojection

View-Independent Volume

Sparse Conv

Fragment Geometry
(sparse TSDF volume)
NeuralRecon

Fragment reconstruction

Feature Unprojection
NeuralRecon
Fragment reconstruction

Input: Posed Images in a fragment
Feature Unprojection
View-Independent Volume
Sparse Conv
Fragment Geometry (sparse TSDF volume)
NeuralRecon

Fragment reconstruction

Input: Posed Images in a fragment
Feature Unprojection
View-Independent Volume
Sparse Conv
Fragment Geometry (sparse TSDF volume)
NeuralRecon
Fragment reconstruction

Why is it better?
Volume-based
Depth-based
NeuralRecon
Fragment reconstruction

Why is it better?

1. Directly predicts the TSDF rather than fusing single-view depth maps
   ==> *learns the shape prior of 3D surfaces, produces locally coherent geometry*
NeuralRecon
Fragment reconstruction

Why is it better?

1. Directly predicts the TSDF rather than fusing single-view depth maps
   ==> learns the shape prior of 3D surfaces, produces locally coherent geometry

2. View-independent volume
   ==> reduces redundant computation, faster
NeuralRecon

Fragment Reconstruction

Fragment Geometry
(sparse TSDF volume)

TSDF Fusion?
NeuralRecon

TSDF Fusion?

Fragment Reconstruction → Fragment Geometry (sparse TSDF volume) → TSDF Fusion → 😞 Too many artifacts!
NeuralRecon

TSDF Fusion?

Fragment Reconstruction → … → ❓ → TSDF Fusion → 

Fragment Geometry  
*(sparse TSDF volume)*

😊 Too many artifacts!

Reason: predicted TSDFs are not consistent between fragments
NeuralRecon
Joint reconstruction and fusion

Input: Posed Images in a fragment

View-Independent Volume

Sparse Conv

Fragment Geometry (sparse TSDF volume)
NeuralRecon
Joint reconstruction and fusion

Input: Posed Images in a fragment
View-Independent Volume

Global Hidden State
GRU Fusion
MLP
Fragment Geometry
(sparse TSDF volume)
NeuralRecon
Joint reconstruction and fusion

Directly fusing the features with GRU
NeuralRecon
Joint reconstruction and fusion

GRU Fusion
Global Hidden State
MLP

Directly fusing the features with GRU

Input: Image Feature Volume
Sparse Conv

Feature:
Extract
Replace

Bounding Volume:
Fragment

Global Hidden State:

Joint reconstruction and fusion

MLP

Directly fusing the features with GRU

Input: Image Feature Volume
Sparse Conv

GRU Fusion
NeuralRecon
Joint reconstruction and fusion

Directly fusing the features with GRU
NeuralRecon
Joint reconstruction and fusion

Directly fusing the features with GRU
NeuralRecon

Joint reconstruction and fusion

Global Hidden State

GRU Fusion

MLP

Replace

Global TSDF Volume
NeuralRecon
Joint reconstruction and fusion
NeuralRecon

Conclusion

View-Independent Volume + Joint TSDF Reconstruction and Fusion → Real-time Coherent

Global Hidden State
GRU Fusion
NeuralRecon

Coarse-to-fine architecture

Input: Fragment Posed Images
NeuralRecon
Coarse-to-fine architecture

Input: Fragment Posed Images
Image Encoder

\{I_t, \xi_t\}_N

\text{Surface Position}
\text{Unoccupied} \quad o \leq \theta
\text{SDF} \quad \begin{cases} 
  o > \theta \\
  x \in (-1, 1) 
\end{cases}
\lambda \quad \text{Truncation Distance}

Output of MLP: \textbf{Occupancy Score} and \textbf{TSDF}
NeuralRecon
Coarse-to-fine architecture

Input: Fragment Posed Images

\{I_t, \xi_t\}_N

Output of MLP: **Occupancy Score** and **TSDF**

Filter by Occupancy Score > 0
NeuralRecon
Coarse-to-fine architecture

Input: Fragment Posed Images

Unprojection

Image Encoder

\{I_t, \xi_t\}_N

Output: Pred. Geo. (Sparse TSDF)

Output of MLP: Occupancy Score and TSDF

Filter by Occupancy Score > 0
NeuralRecon
Coarse-to-fine architecture

End-to-end learnable volumetric reconstruction system

Input: Fragment Posed Images
Image Encoder

Output of MLP: Occupancy Score and SDF

Output: Pred. Geo. (Sparse TSDF)
NeuralRecon

Training

ScanNet dataset

Contains 2.5M RGB images captured in more than 1500 scans annotated with 3D camera poses and surface reconstructions

Binary cross-entropy (BCE) and L1 loss are used for training
Experiments

Qualitative results: office 1

- Ours (30ms)
- COLMAP (2076ms)
- DeepV2D (347ms)
- Ground Truth
- CNMNet (80ms)
- Atlas (292ms)
Experiments

Qualitative results: office 2

Ours (30ms)  COLMAP (2076ms)  DeepV2D (347ms)

Ground Truth  CNMNet (80ms)  Atlas (292ms)
Experiments

Qualitative results: Comparison with Atlas on a **large** scene (30m x 10m)

Ours
Max GPU Memory: 3.29GB

Atlas
Max GPU Memory: >24GB (OOM)

The reconstruction is incomplete due to out of memory (OOM) error on the remaining sequence.
Experiments
Quantitative results

Real-time methods:  ▼ MVDepthNet  ▶ GPMVS

<table>
<thead>
<tr>
<th>Speed</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>😞</td>
<td>😞</td>
</tr>
</tbody>
</table>

Graph showing the comparison of F-score with Time [ms] for MVDepthNet and GPMVS.
Experiments
Quantitative results

![Graph showing F-score vs. Time for various methods.]

**Real-time methods:**
- MVDepthNet
- GPMVS

**Multiple View Stereo methods:**
- DPSNet
- COLMAP
- Atlas

<table>
<thead>
<tr>
<th>Method</th>
<th>Speed</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVDepthNet</td>
<td>😞</td>
<td>😞</td>
</tr>
<tr>
<td>GPMVS</td>
<td>😞</td>
<td>😞</td>
</tr>
<tr>
<td>DPSNet</td>
<td>😞</td>
<td>😞</td>
</tr>
<tr>
<td>COLMAP</td>
<td>😞</td>
<td>😞</td>
</tr>
<tr>
<td>Atlas</td>
<td>😞</td>
<td>😞</td>
</tr>
</tbody>
</table>
Experiments
Quantitive results

Real-time methods:
- MVDepthNet
- GPMVS

Multiple View Stereo methods:
- DPSNet
- COLMAP
- Atlas
- Ours

<table>
<thead>
<tr>
<th>Method</th>
<th>Time [ms]</th>
<th>F-score</th>
<th>Speed</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVDepthNet</td>
<td>200</td>
<td>0.55</td>
<td>😞</td>
<td>😞</td>
</tr>
<tr>
<td>GPMVS</td>
<td>500</td>
<td>0.50</td>
<td>😞</td>
<td>😞</td>
</tr>
<tr>
<td>Ours</td>
<td>1000</td>
<td>0.45</td>
<td>😃</td>
<td>😃</td>
</tr>
<tr>
<td>DPSNet</td>
<td>2000</td>
<td>0.40</td>
<td>😃</td>
<td>😃</td>
</tr>
<tr>
<td>COLMAP</td>
<td></td>
<td></td>
<td>😞</td>
<td>😞</td>
</tr>
<tr>
<td>Atlas</td>
<td></td>
<td></td>
<td>😞</td>
<td>😞</td>
</tr>
</tbody>
</table>
Experiments
Quantitative results

<table>
<thead>
<tr>
<th>Method</th>
<th>Time [ms]</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVDepthNet</td>
<td>250</td>
<td>0.329</td>
</tr>
<tr>
<td>GPMVS</td>
<td>750</td>
<td>0.562</td>
</tr>
<tr>
<td>COLMAP</td>
<td>1250</td>
<td>0.30</td>
</tr>
<tr>
<td>Atlas</td>
<td>2000</td>
<td>0.35</td>
</tr>
<tr>
<td>DPSNet</td>
<td>500</td>
<td>0.55</td>
</tr>
<tr>
<td>Ours</td>
<td>1000</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Real-time methods:  
- MVDepthNet  
- GPMVS

Multiple View Stereo methods:  
- DPSNet  
- COLMAP  
- Atlas  
- Ours

Speed:  
- 😞
- 😞
- 😃
- 😃

Accuracy:  
- 😞
- 😞
- 😃
- 😃
Experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>Time [ms]</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVDepthNet</td>
<td>30</td>
<td>Better</td>
</tr>
<tr>
<td>GPMVS</td>
<td>2076</td>
<td>Better</td>
</tr>
</tbody>
</table>

Real-time methods:
- MVDepthNet
- GPMVS

Multiple View Stereo methods:
- DPSNet
- COLMAP
- Atlas
- Ours

Speed: 😞
Accuracy: 😞
Demo

Indoor scene at our office
Demo
Indoor scene with extremely low texture
Demo

Generalization to outdoor scenes
Demo
AR Demo
NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video

Jiaming Sun* Yiming Xie* Linghao Chen Xiaowei Zhou Hujun Bao
CVPR 2021 (Oral)

Project page: https://zju3dv.github.io/neuralrecon/
Code: https://zju3dv.github.io/NeuralRecon/
Thanks for watching!

Q&A