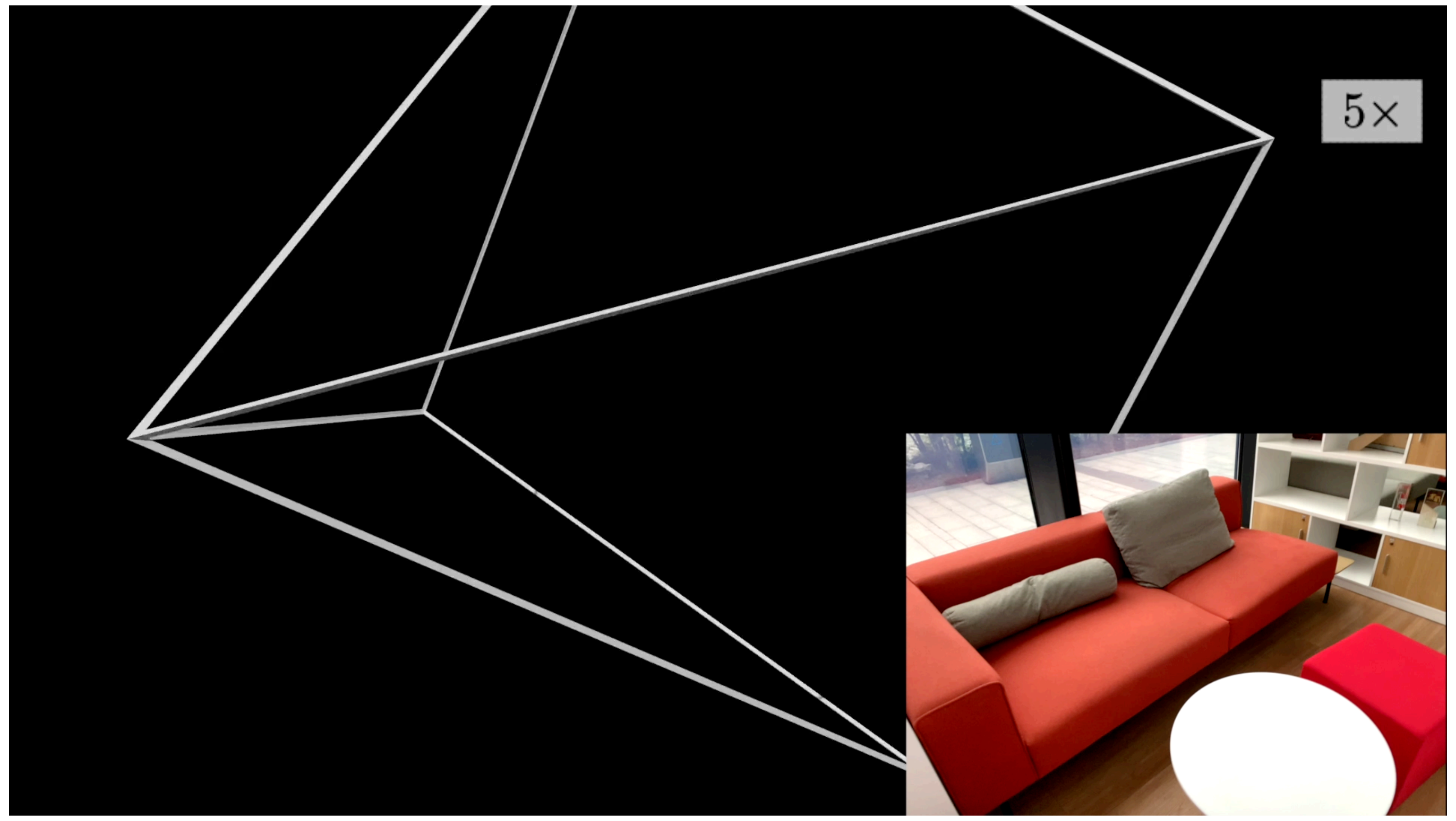
NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video

Jiaming Sun* Yiming Xie* Linghao Chen Xiaowei Zhou Hujun Bao

CVPR 2021 (Oral and Best Paper Candidate)

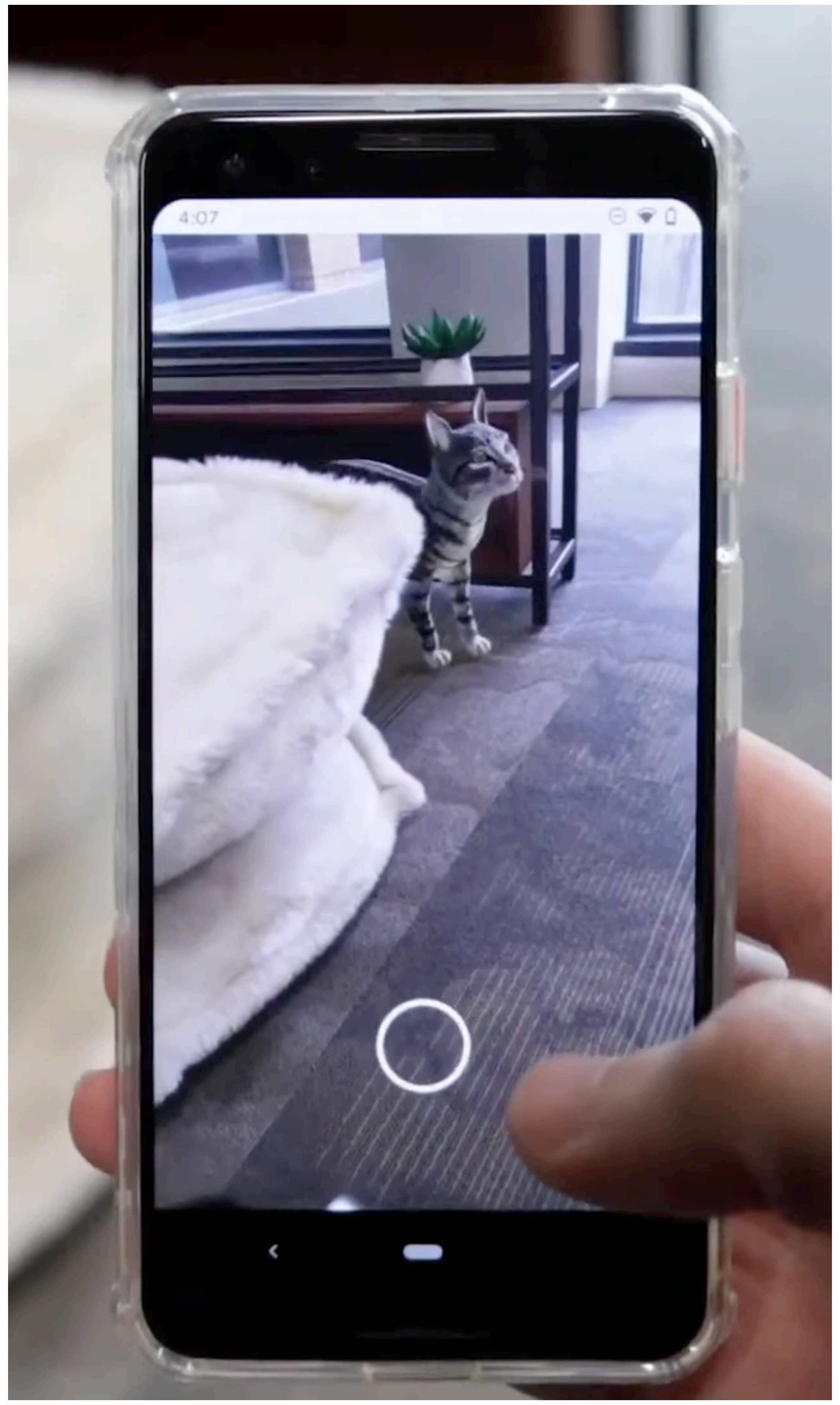
* equal contribution



Motivation

~~3D Geometry~~ is crucial for immersive AR effects

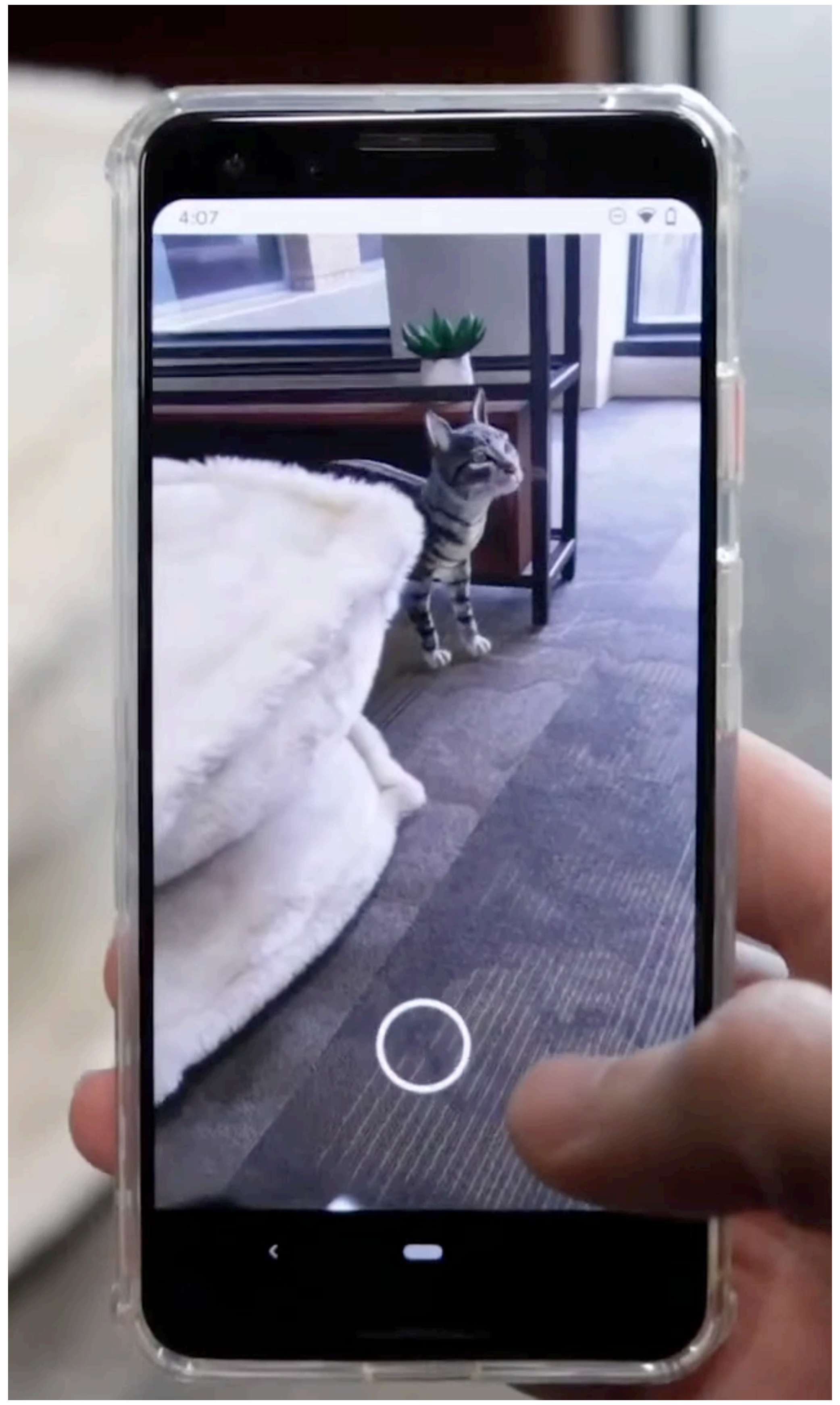
3D Reconstruction



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

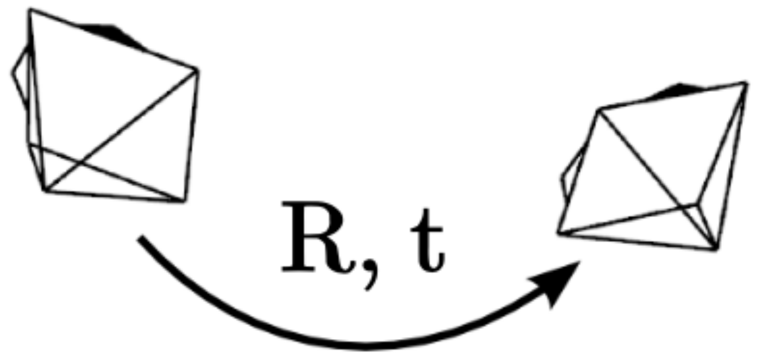
😃 Accurate depth measurement

🙁 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

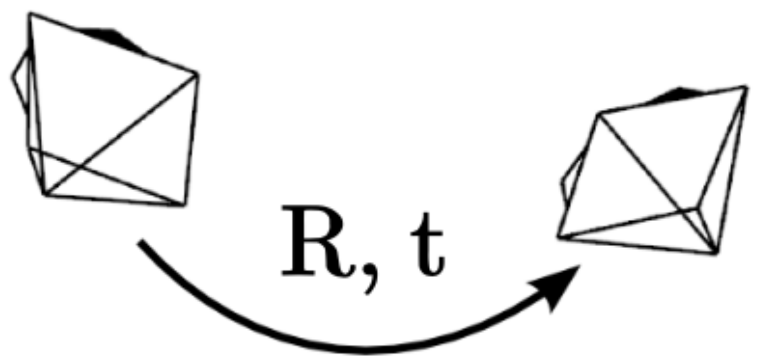
😃 Accurate depth measurement

🙁 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

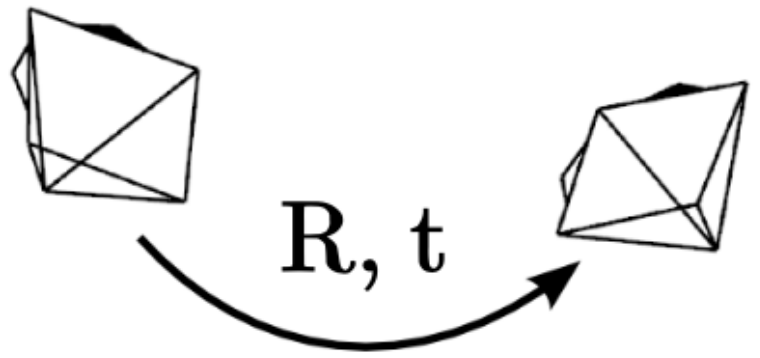
😃 Accurate depth measurement

🙁 Takes a lot of energy

🙁 Only available to a few high-end products

With a monocular camera

+



😃 Immediately available to many phones

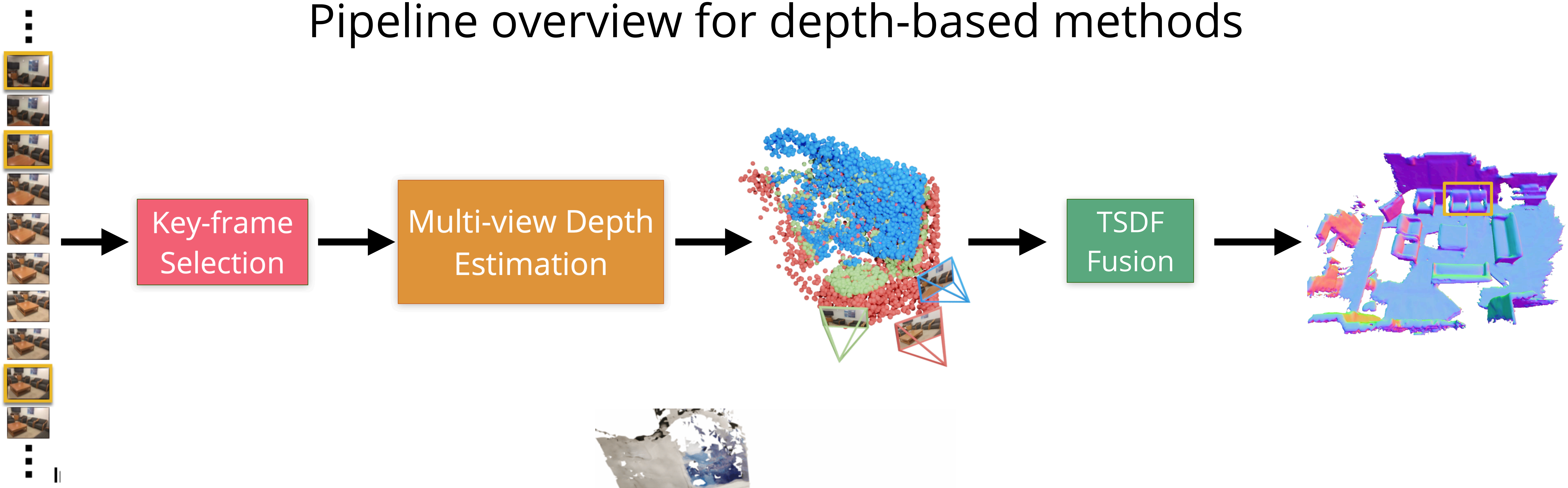
🙁 Not as accurate as depth sensors

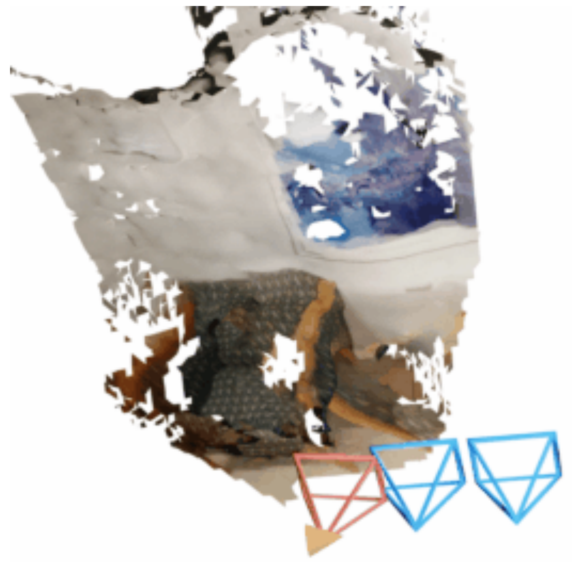
🙁 Not as fast as depth sensors

Credit: RealSense, Azure Kinect, LiDAR on iPad Pro

Motivation

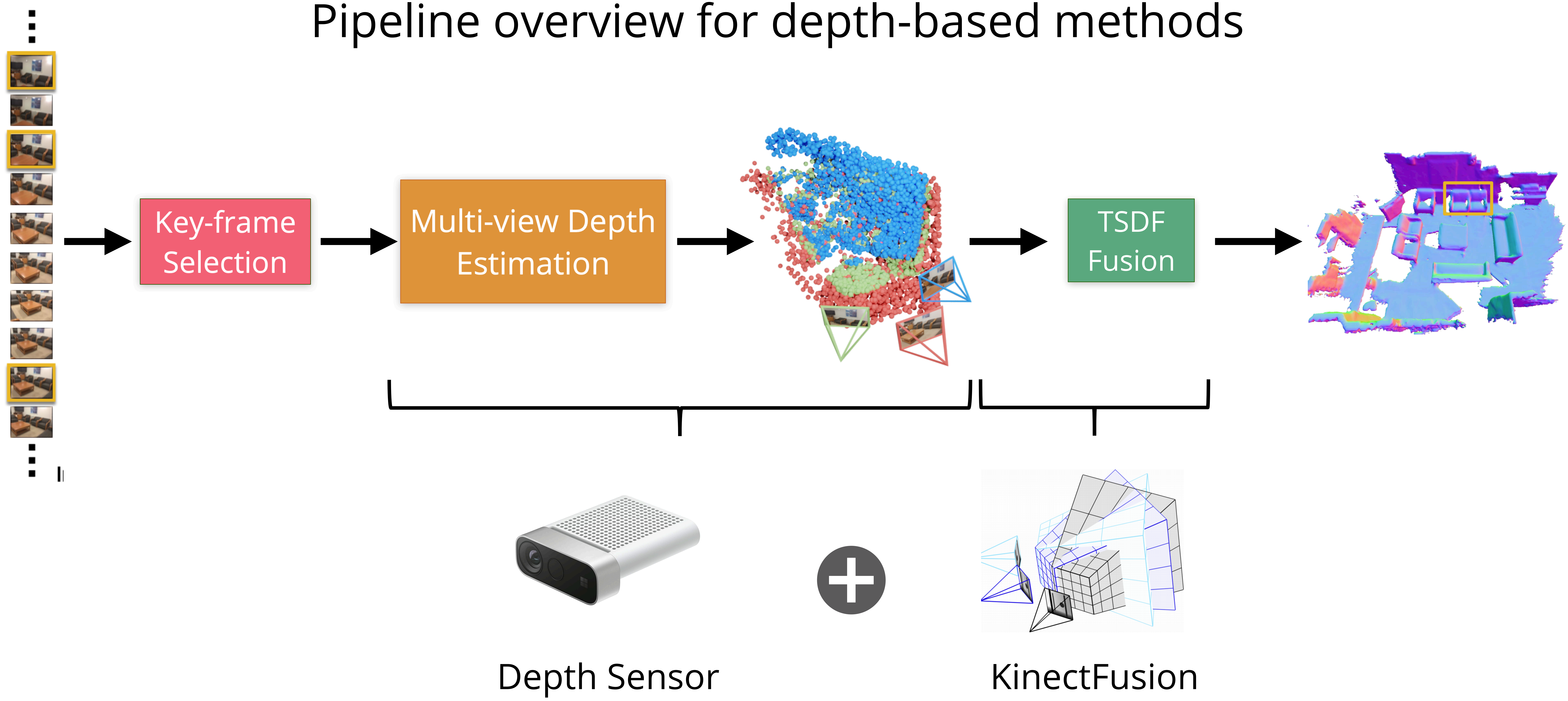
Pipeline overview for depth-based methods





Motivation

Pipeline overview for depth-based methods

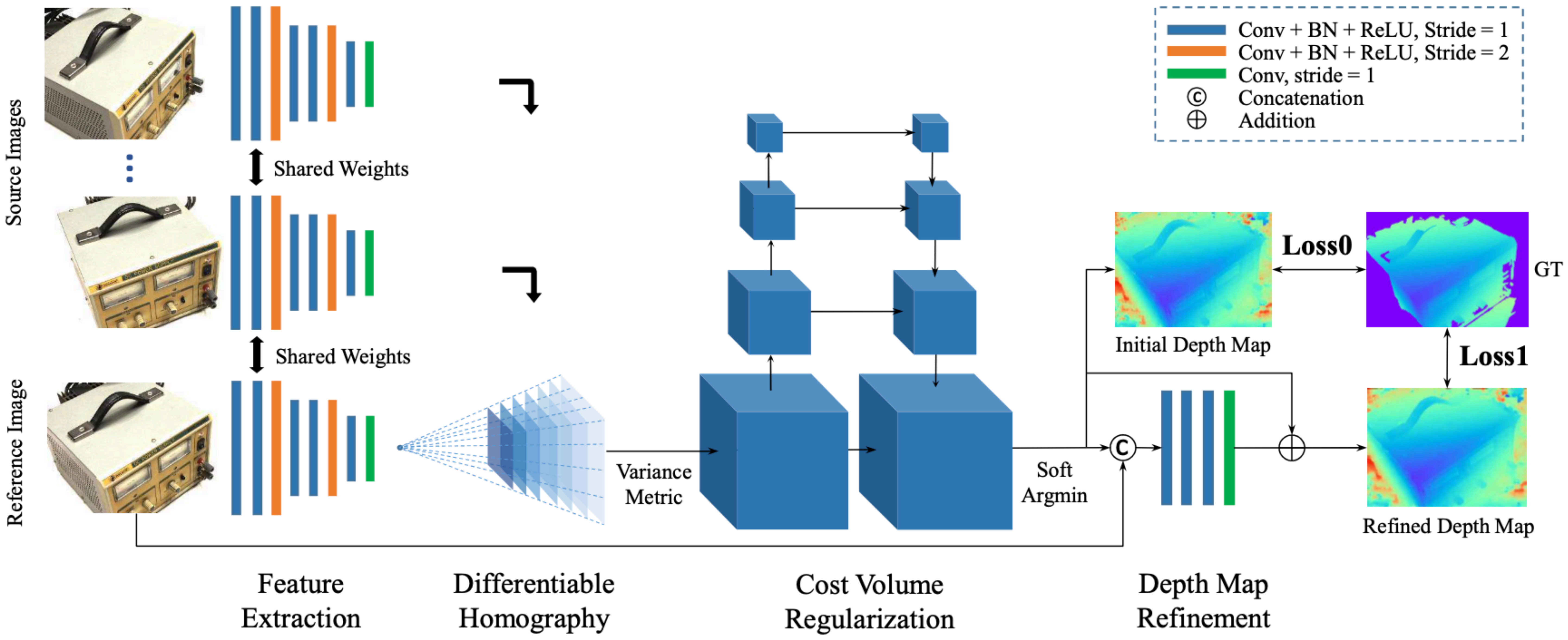


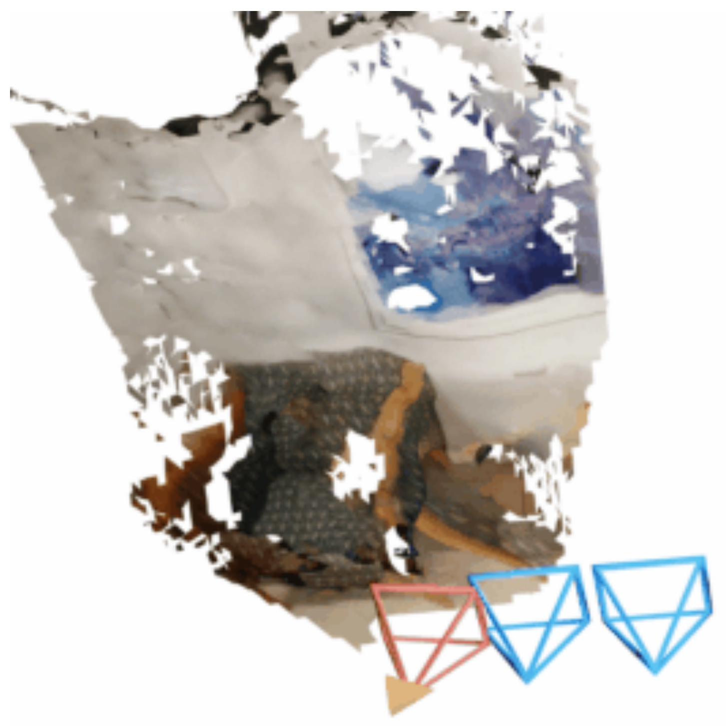
Newcombe et al., "**KinectFusion: Real-Time Dense Surface Mapping and Tracking**", *UIST 2011*

Motivation

MVSNet: Depth Inference for Unstructured Multi-view Stereo

Yao Yao[1], Zixin Luo[1], Shiwei Li[1], Tian Fang[2], and Long Quan[1]





Motivation

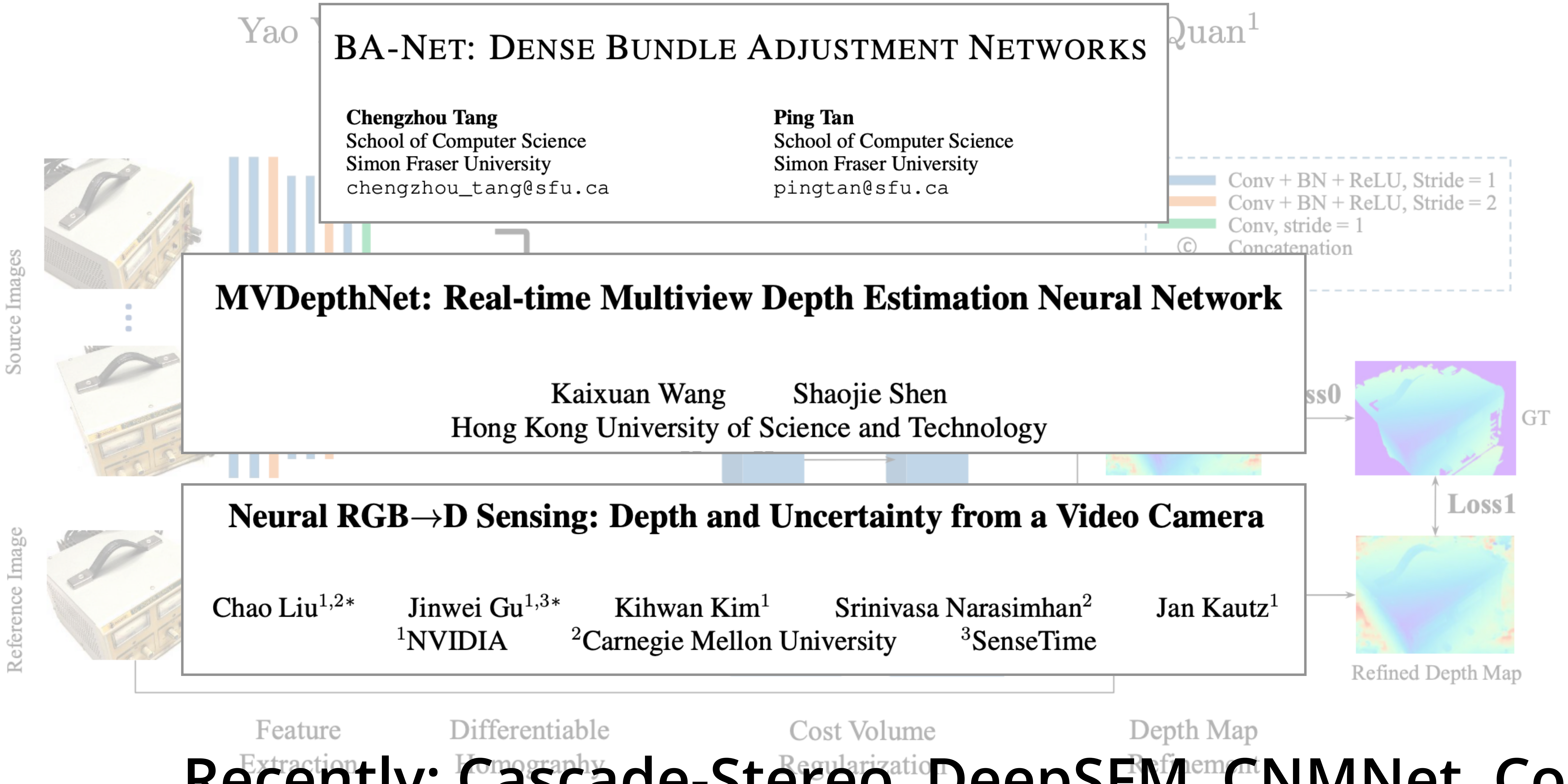
DeepTAM: Deep Tracking and Mapping

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
Simon Fraser University
chengzhou_tang@sfu.ca

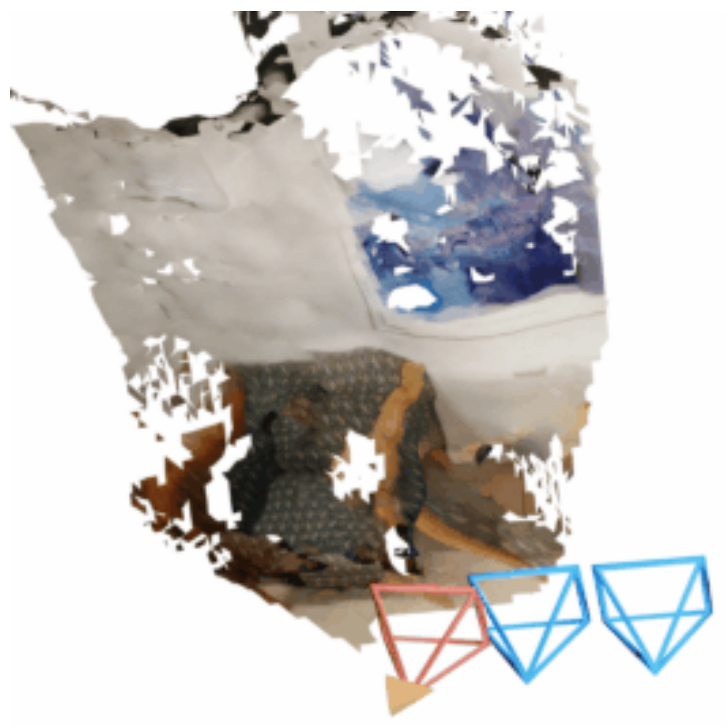
Ping Tan
School of Computer Science
Simon Fraser University
pingtan@sfu.ca

MVDepthNet: 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 Liu[1,2*] Jinwei Gu[1,3*] Kihwan Kim[1] Srinivasa Narasimhan[2] Jan Kautz[1]
[1]NVIDIA [2]Carnegie Mellon University [3]SenseTime

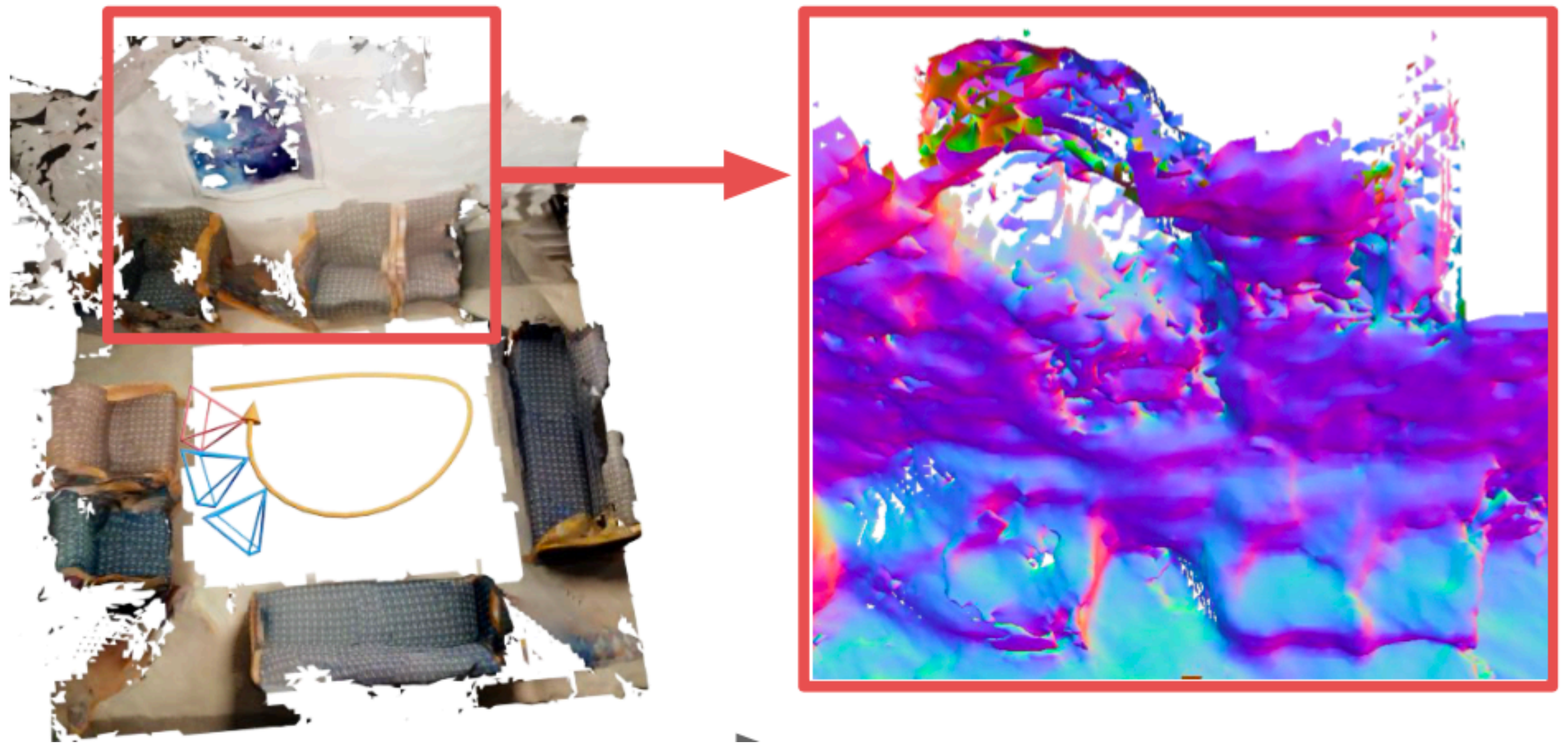


Recently: Cascade-Stereo, DeepSFM, CNMNet, Consistent Depth…

Motivation

Depth-based methods v.s. NeuralRecon

Depth-based methods

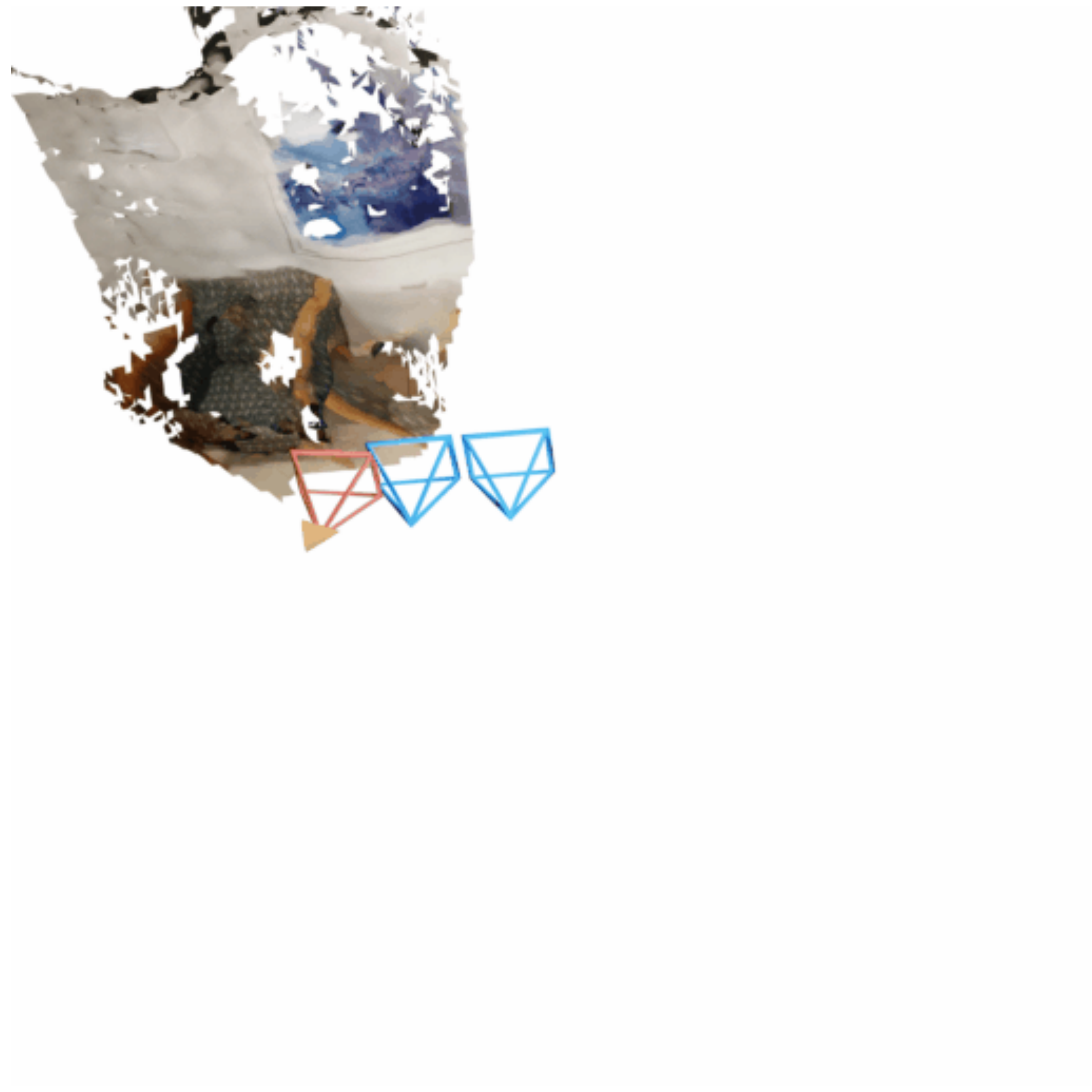


🙁 Either layered or scattered results

Motivation

Depth-based methods v.s. NeuralRecon

Depth-based methods



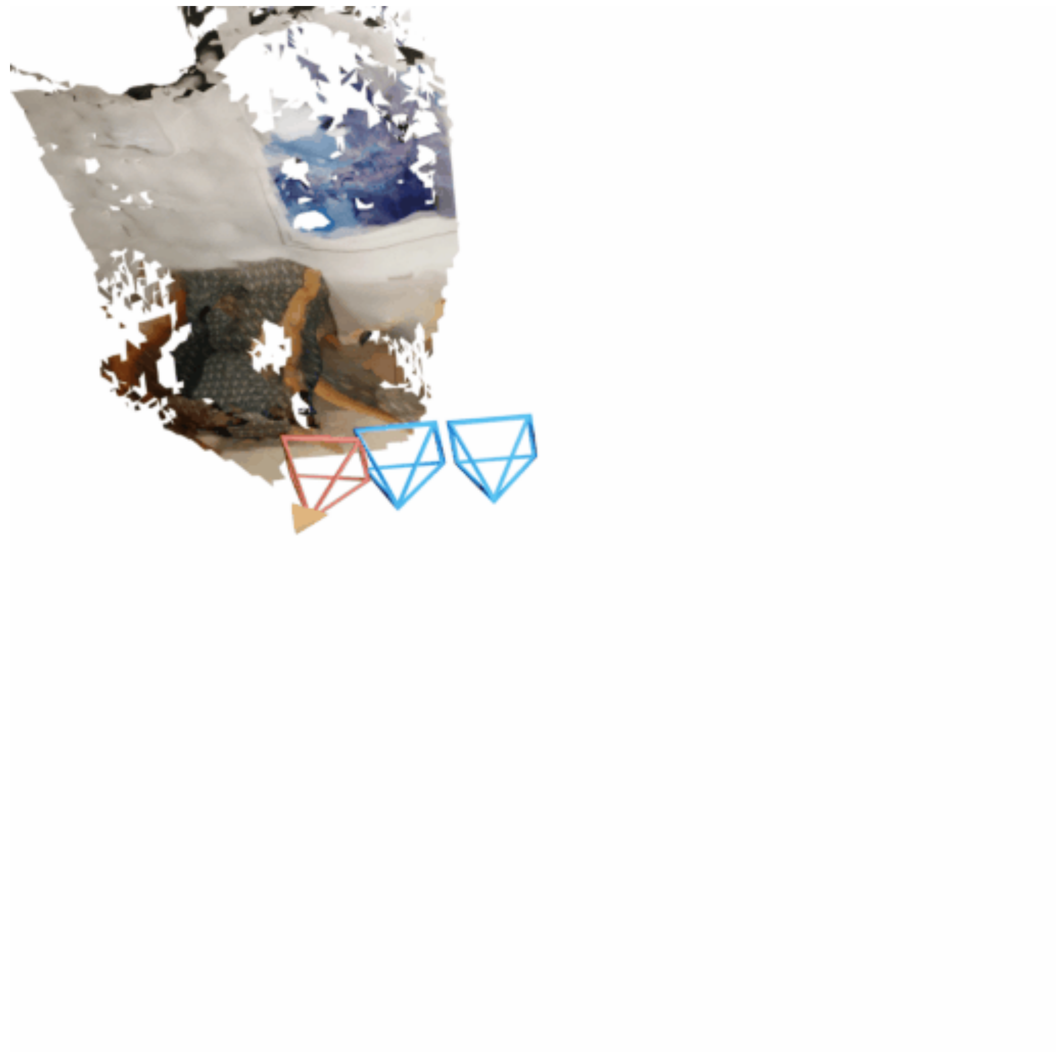
🙁 Either layered or scattered results

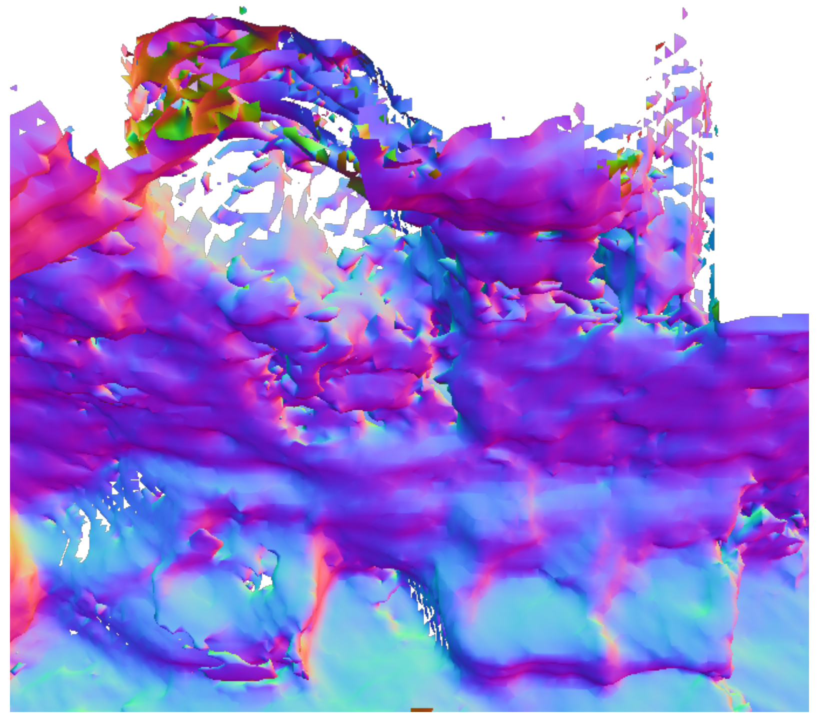
🙁 Redundant computation

Motivation

Depth-based methods v.s. NeuralRecon

Depth-based methods

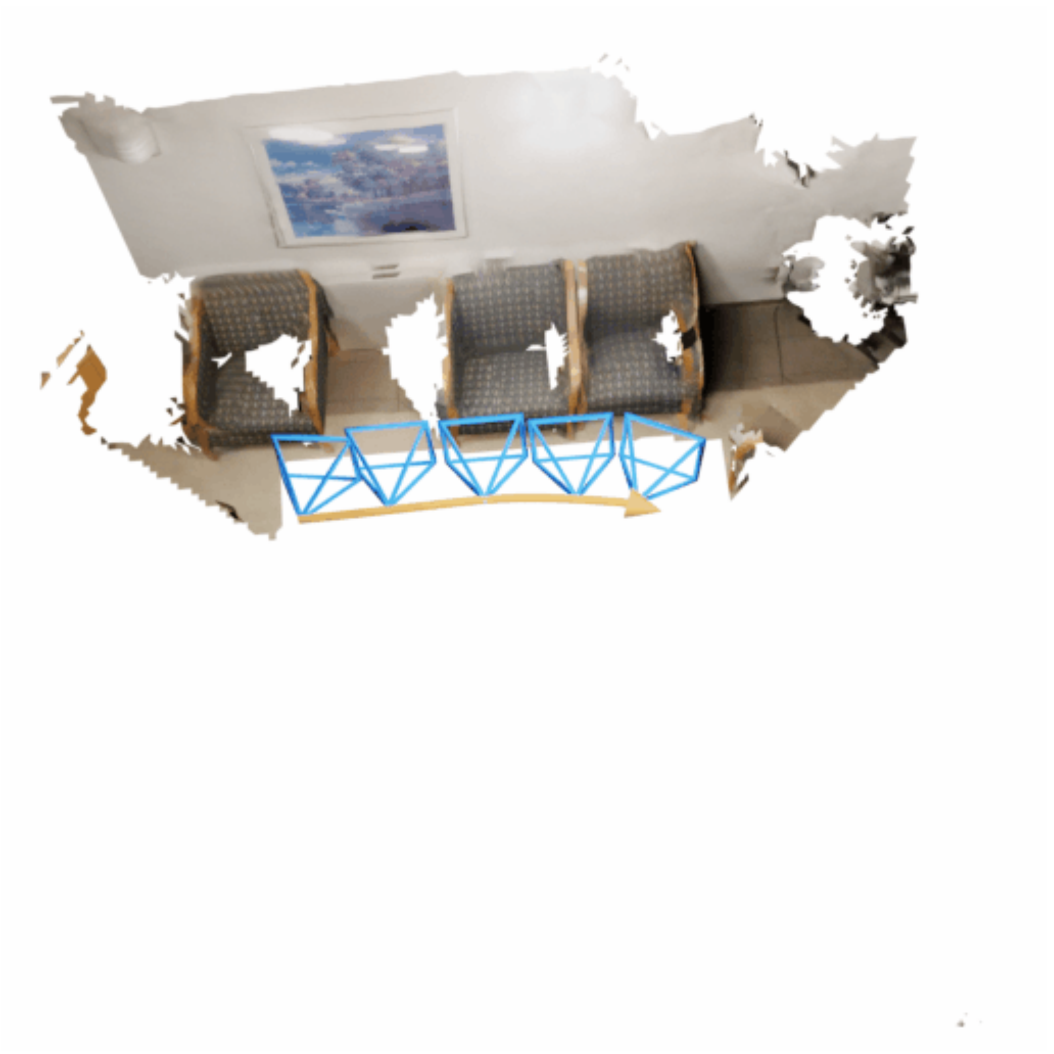


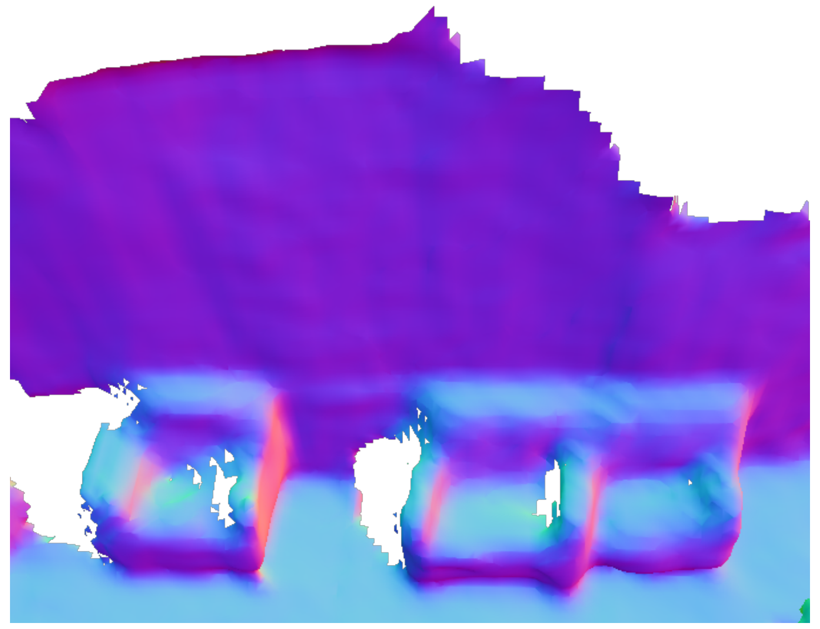


🙁 Either layered or scattered results

🙁 Redundant computation

Our solution: NeuralRecon





😃 Reconstruct local surfaces directly in TSDF

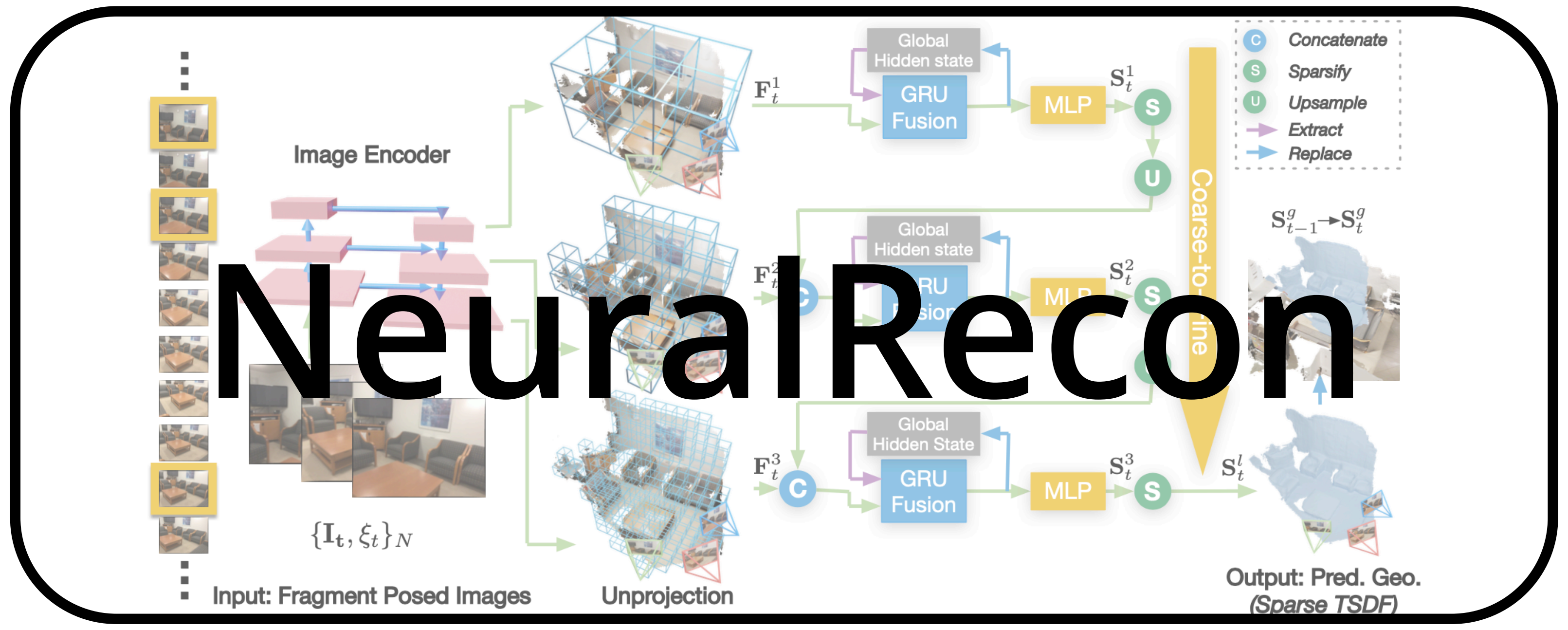
😃 Joint fragment reconstruction and fusion

😃 Better quality and faster speed

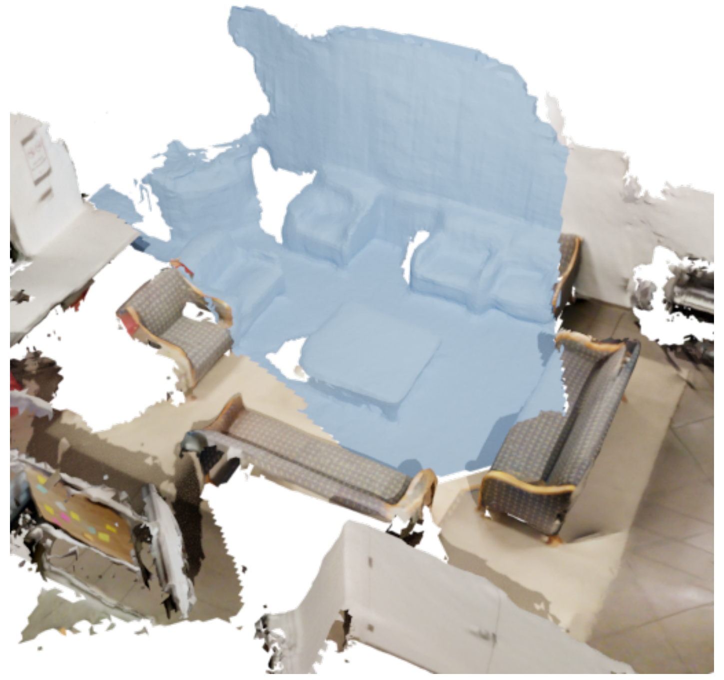
NeuralRecon

Overview

Input:
Posed Images



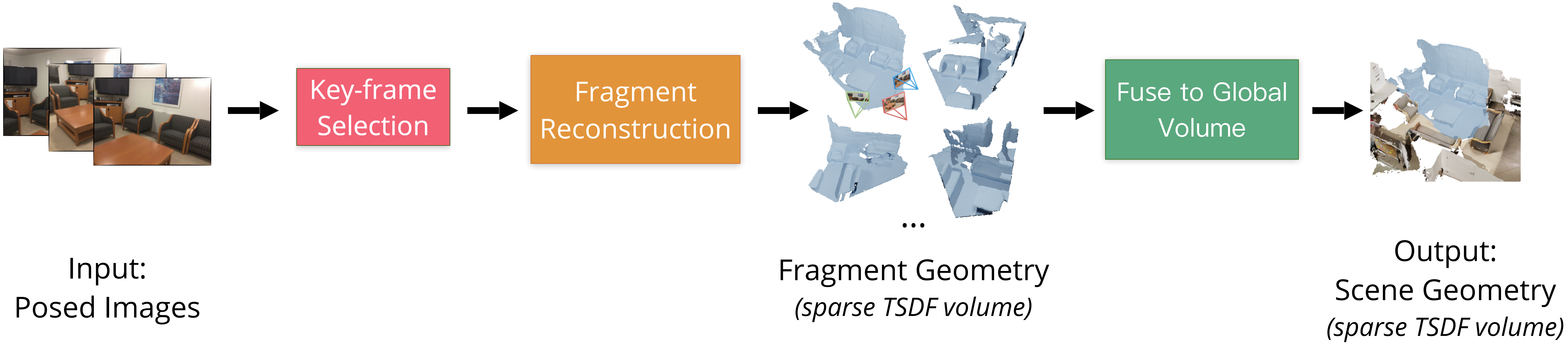
End-to-End System



Output:
Scene Geometry
(sparse TSDF volume)

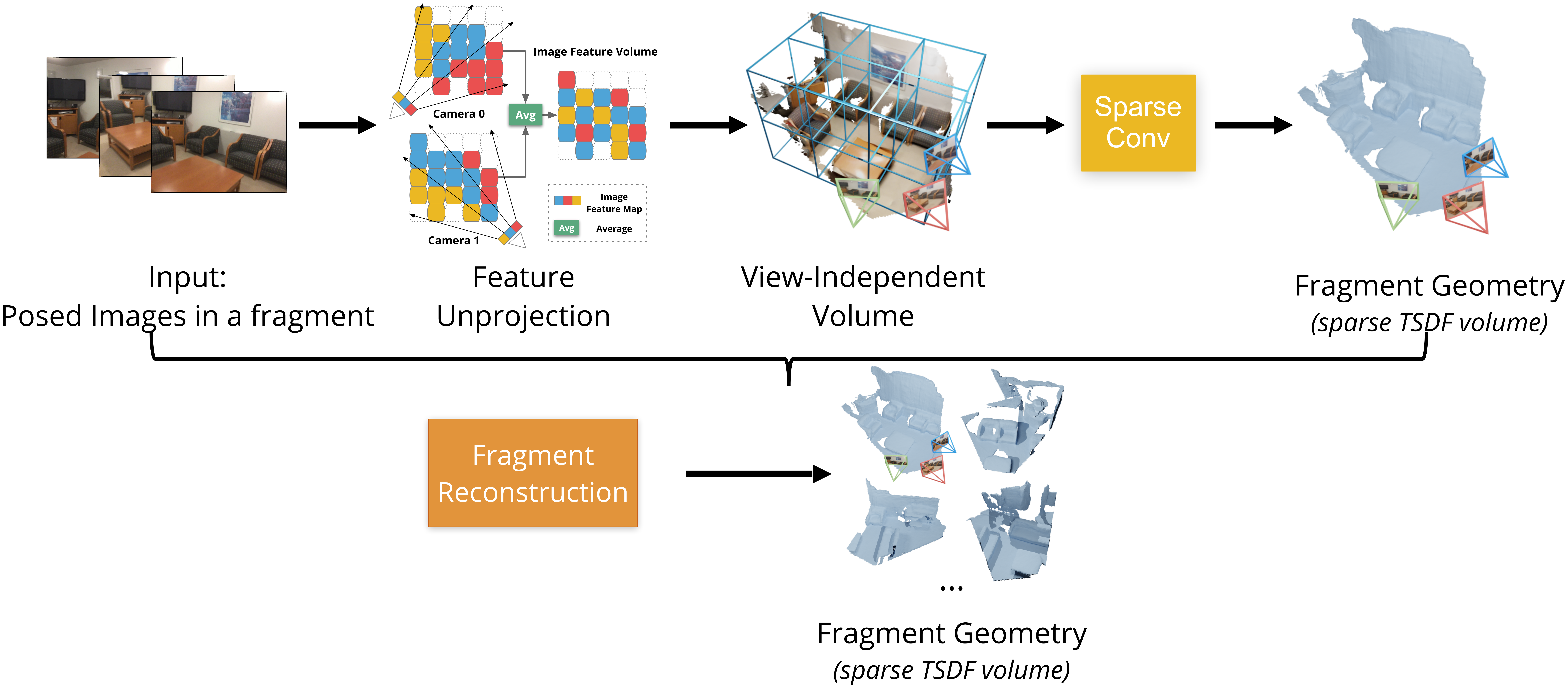
NeuralRecon

Overview



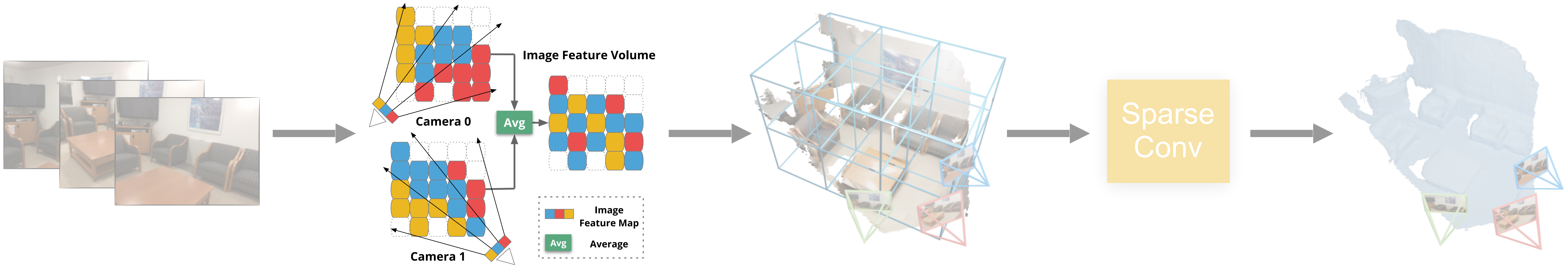
NeuralRecon

Fragment reconstruction



NeuralRecon

Fragment reconstruction



Input:
Posed Images in a fragment

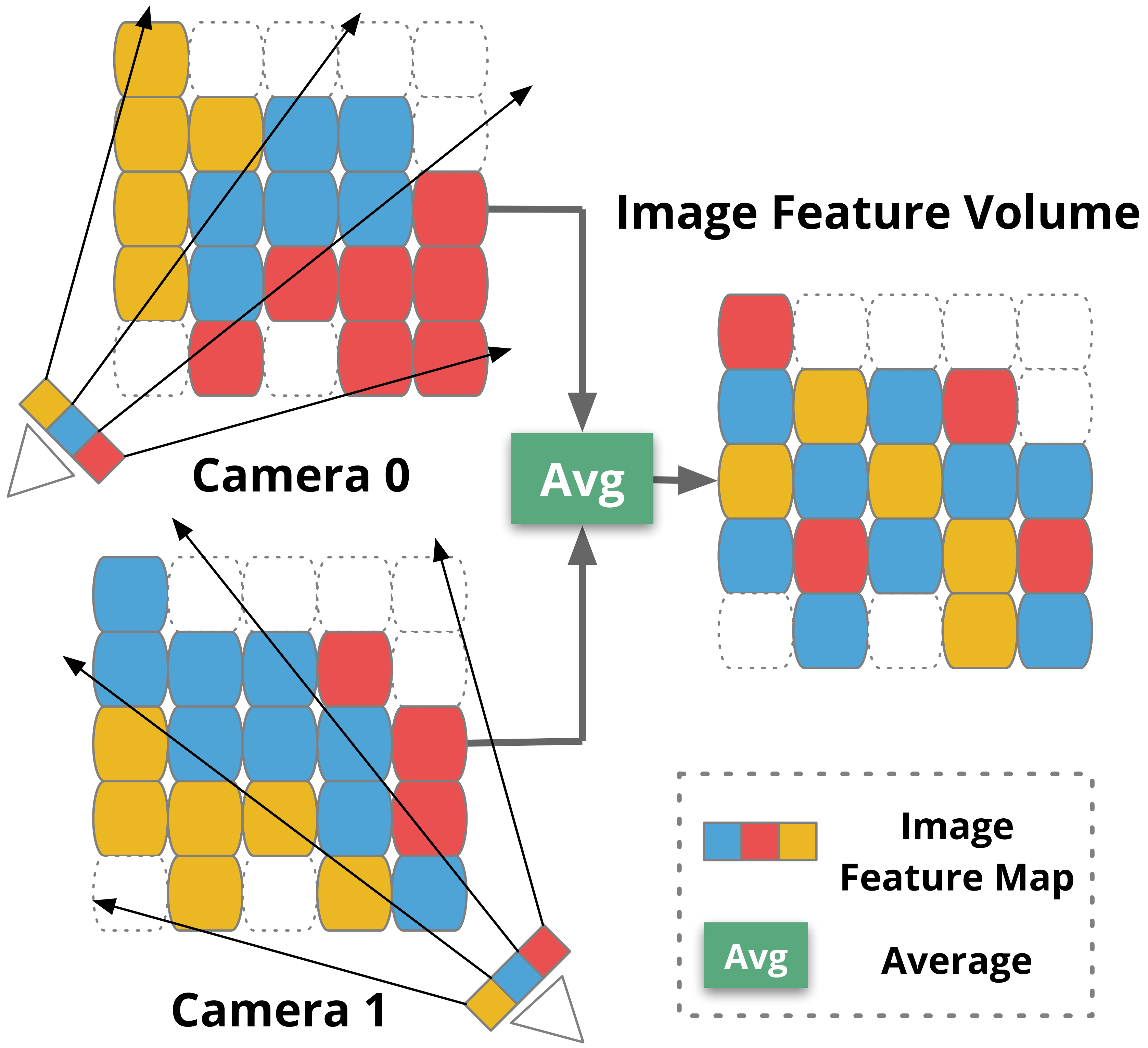
Feature
Unprojection

View-Independent
Volume

Fragment Geometry
(sparse TSDF volume)

NeuralRecon

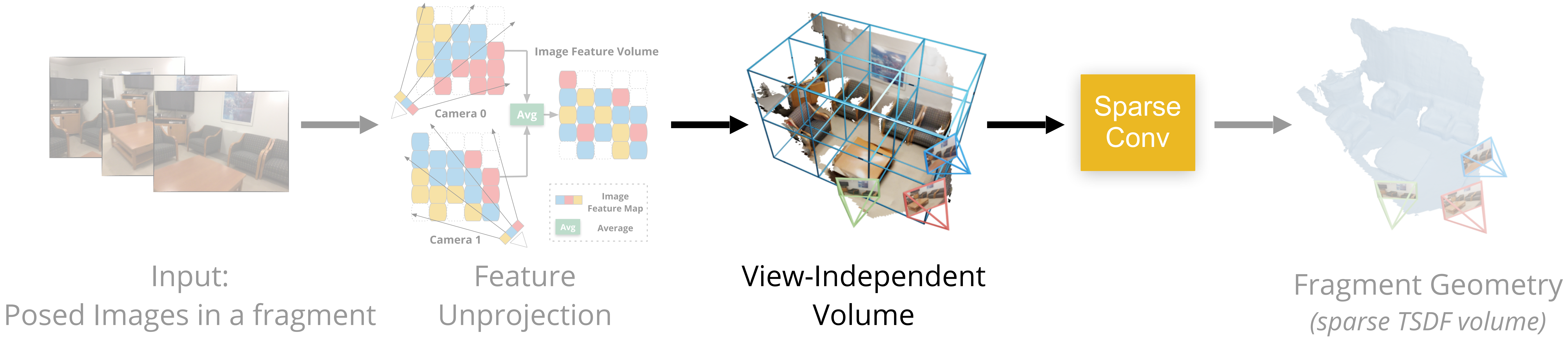
Fragment reconstruction



Feature
Unprojection

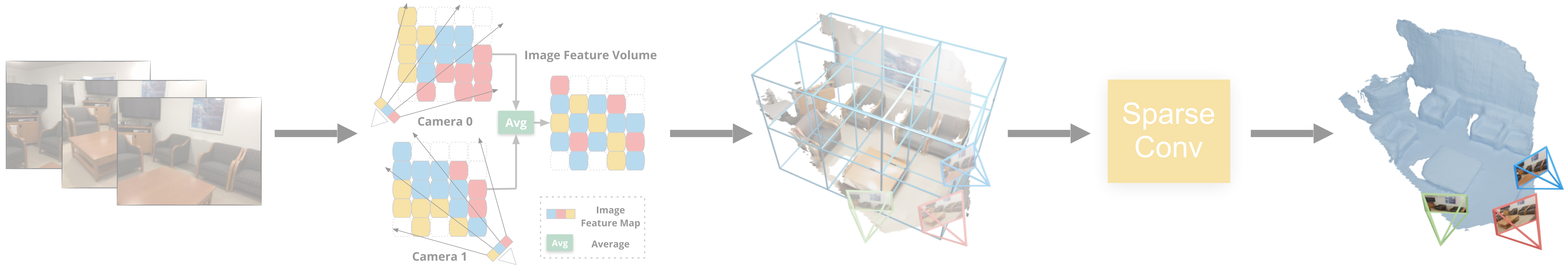
NeuralRecon

Fragment reconstruction



NeuralRecon

Fragment reconstruction



Input:
Posed Images in a fragment

Feature
Unprojection

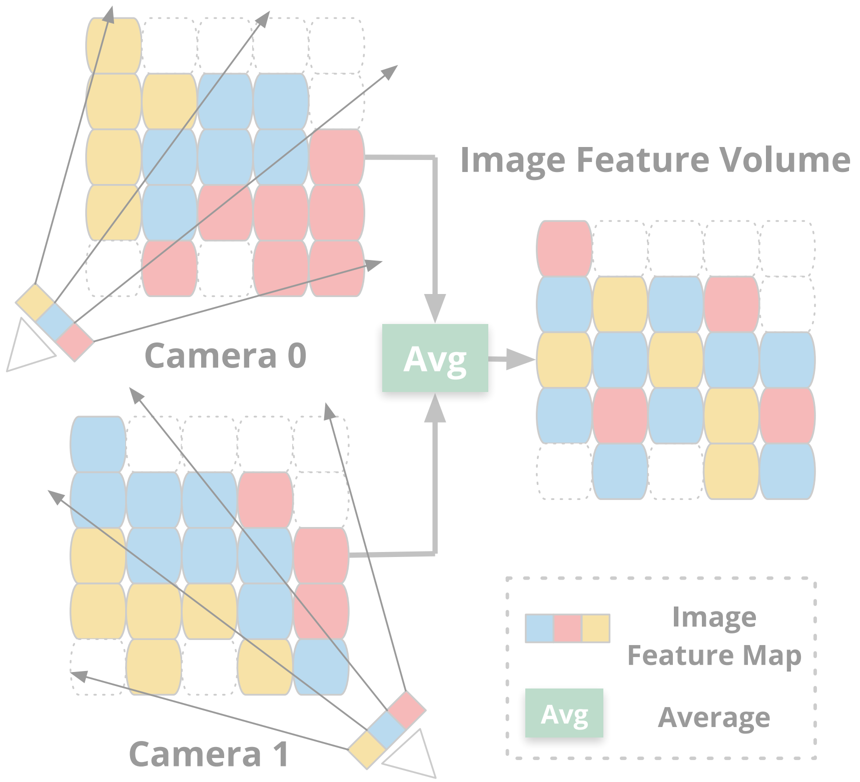
View-Independent
Volume

Fragment Geometry
(sparse TSDF volume)

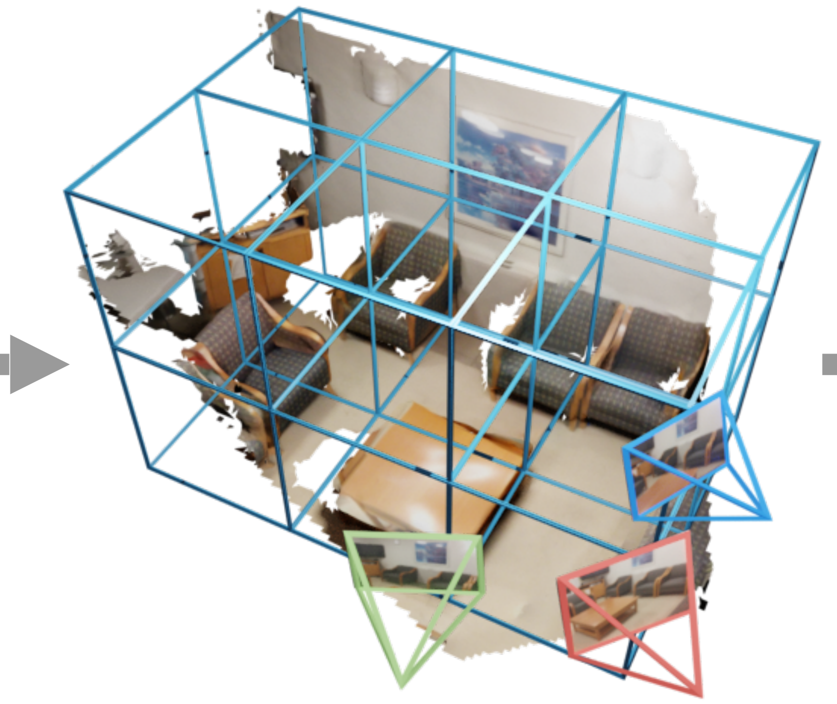
NeuralRecon

Fragment reconstruction

Input:
Posed Images in a fragment



Feature
Unprojection

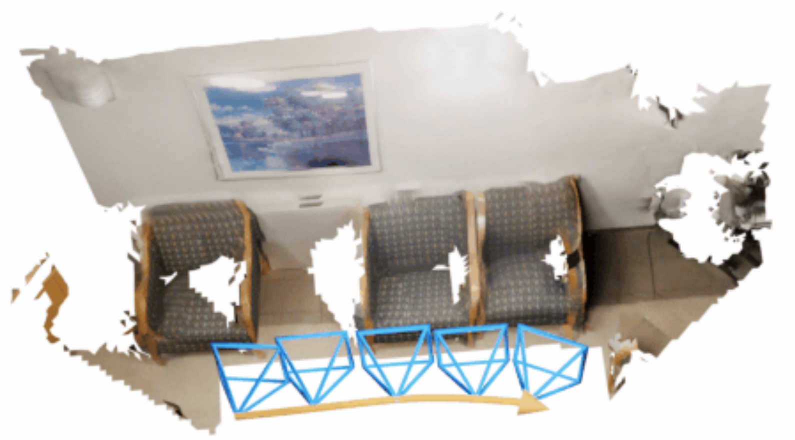


View-Independent
Volume

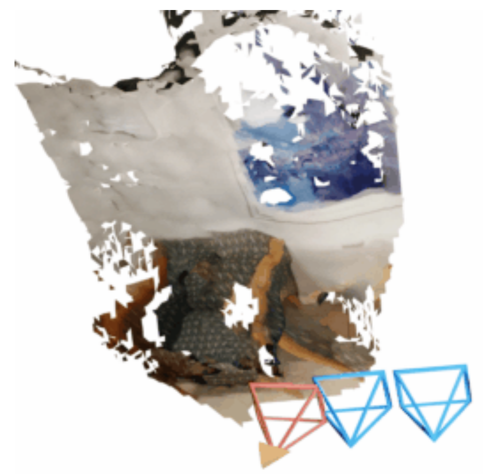
Sparse
Conv

Fragment Geometry
(sparse TSDF volume)

Why is it better?



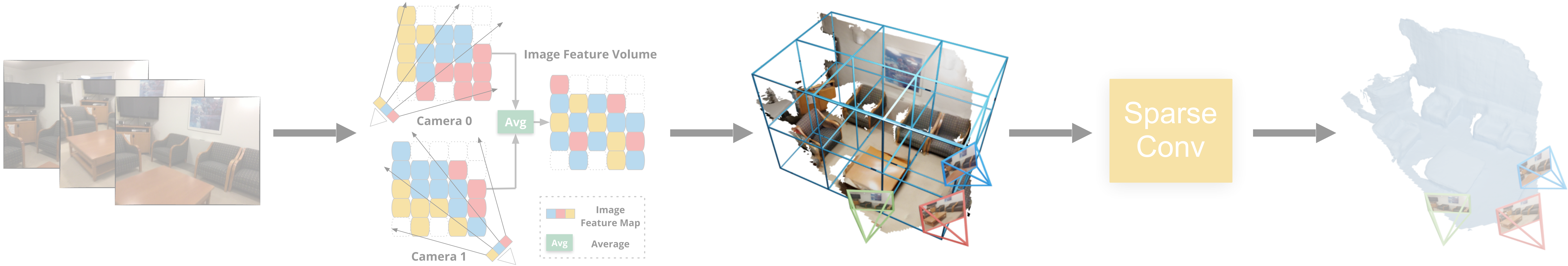
Volume-based



Depth-based

NeuralRecon

Fragment reconstruction



Input:
Posed Images in a fragment

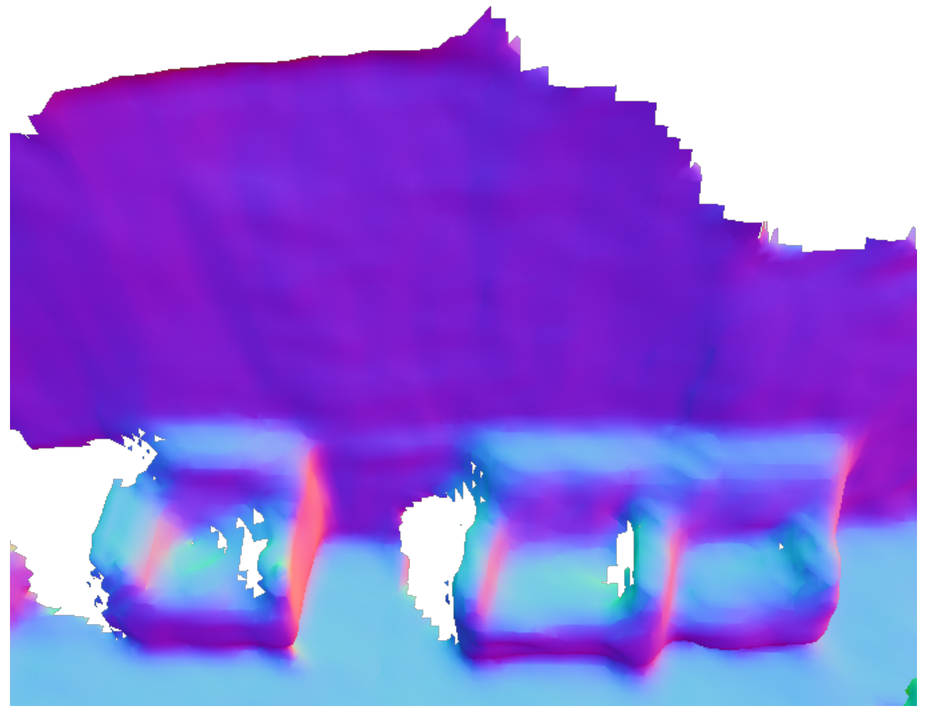
Feature
Unprojection

View-Independent
Volume

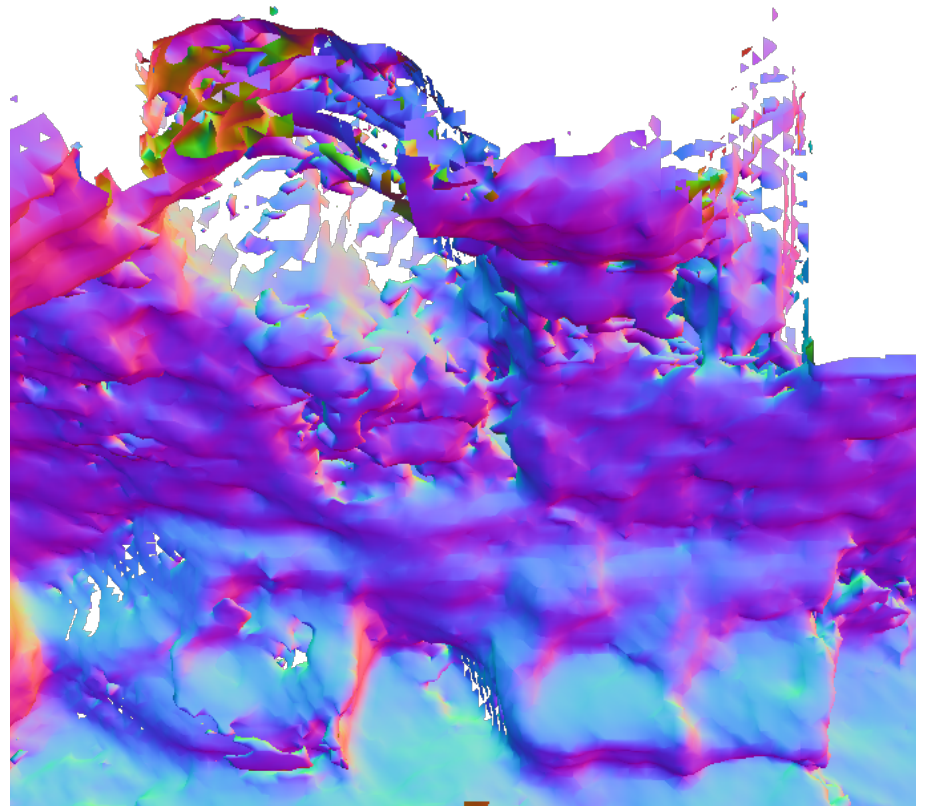
Fragment Geometry
(sparse TSDF volume)

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*



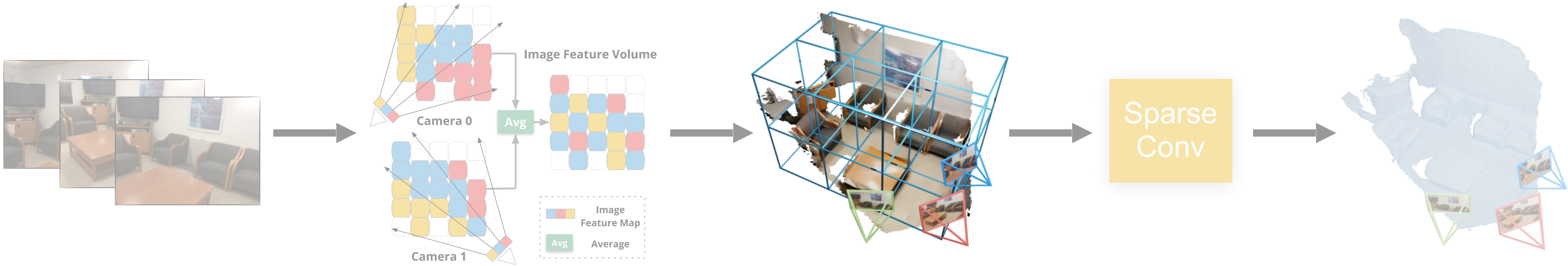
Volume-based



Depth-based

NeuralRecon

Fragment reconstruction



Input:
Posed Images in a fragment

Feature
Unprojection

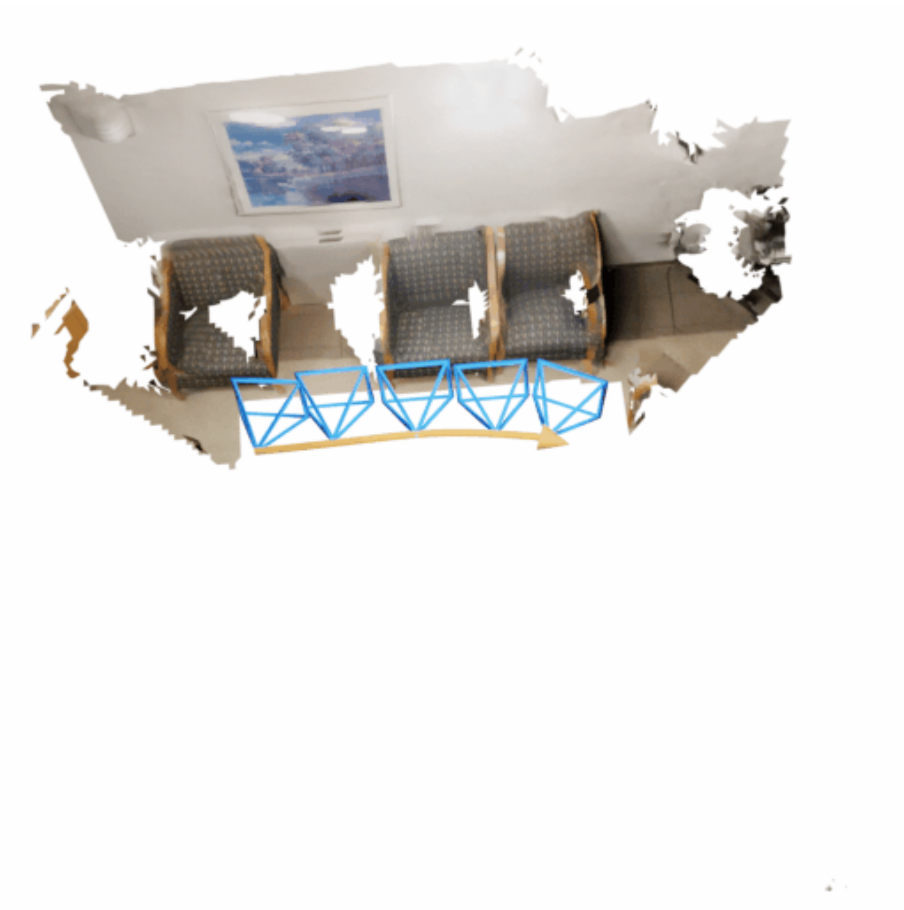
View-Independent
Volume

Fragment Geometry
(sparse TSDF volume)

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*



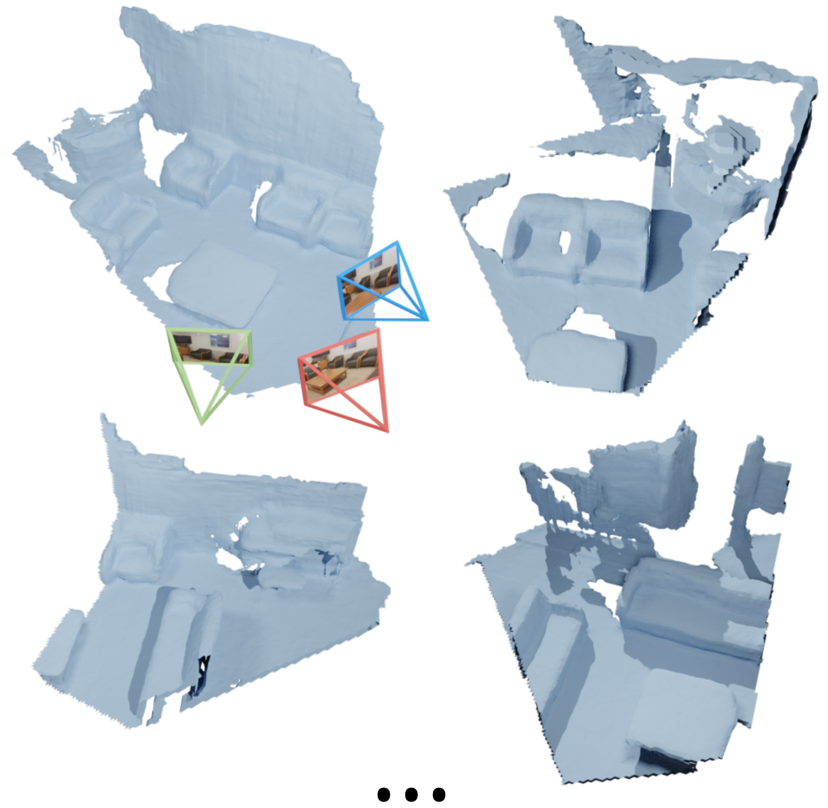
Volume-based

Depth-based

NeuralRecon

TSDF Fusion?

Fragment Reconstruction



...

Fragment Geometry

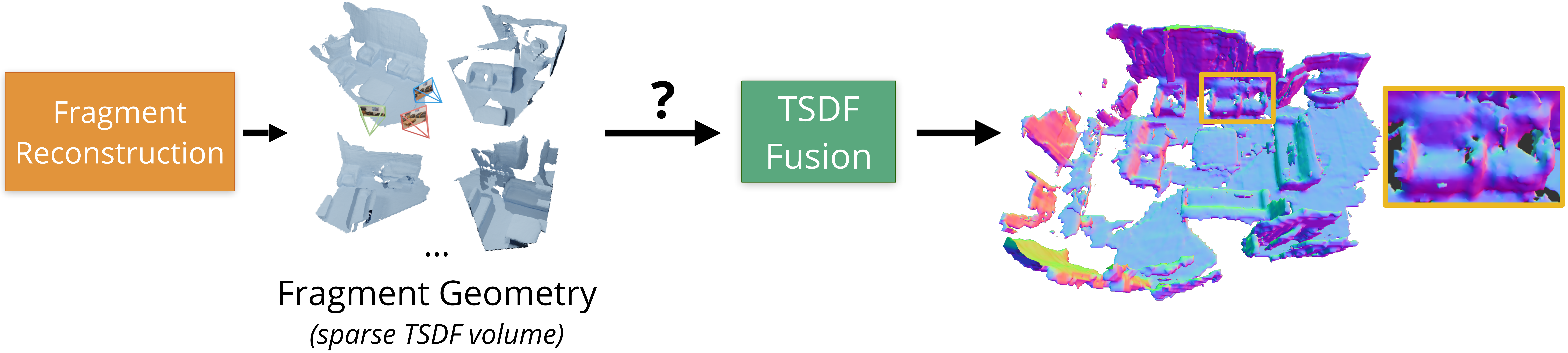
(sparse TSDF volume)

?

TSDF Fusion

NeuralRecon

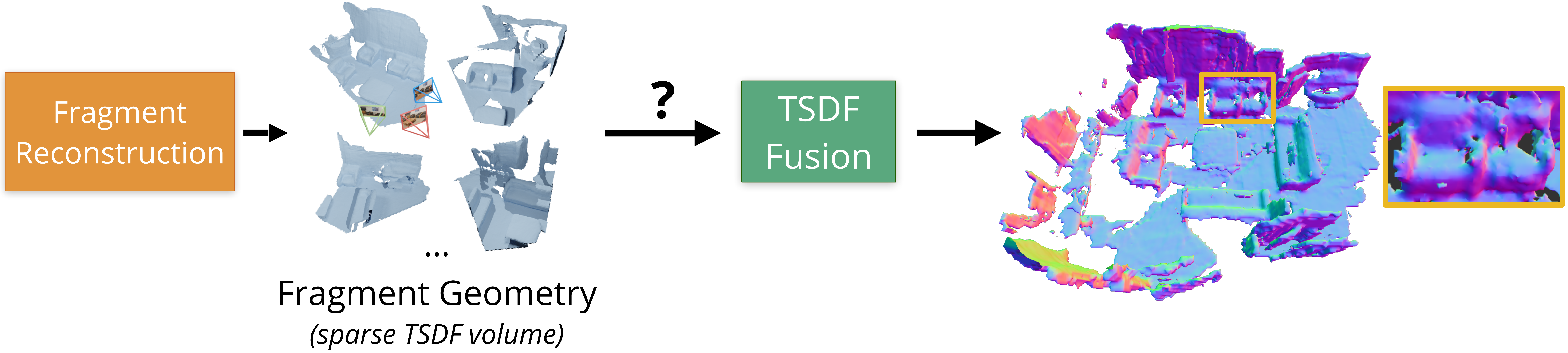
TSDF Fusion?



🙁 **Too many artifacts!**

NeuralRecon

TSDF Fusion?



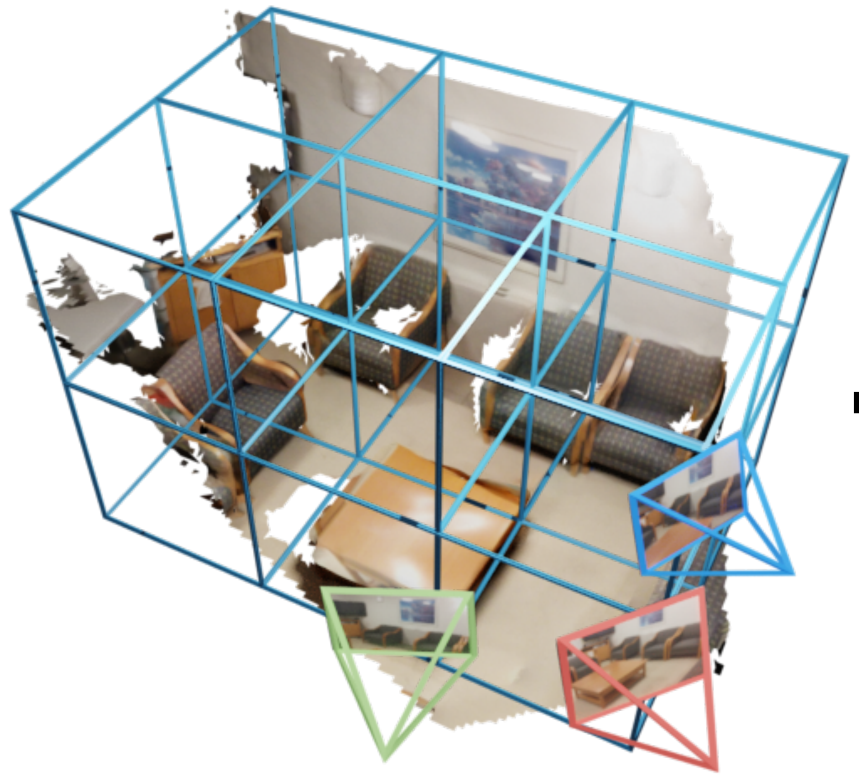
🙁 **Too many artifacts!**

Reason: predicted TSDFs are not consistent between fragments

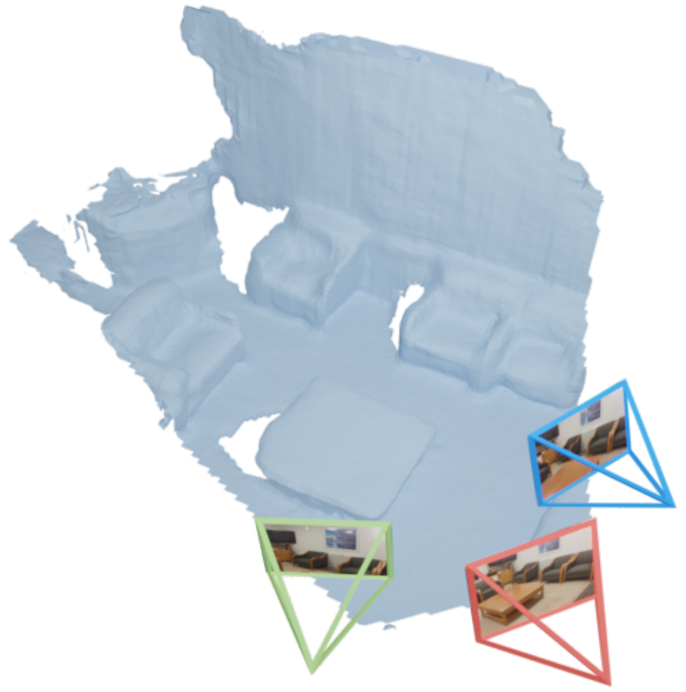
NeuralRecon

Joint reconstruction and fusion

Input:
Posed Images in a fragment



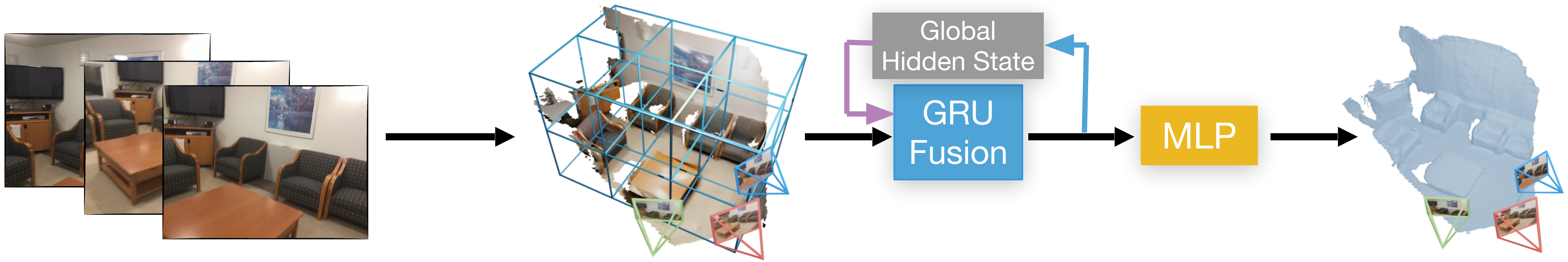
View-Independent
Volume



Fragment Geometry
(sparse TSDF volume)

NeuralRecon

Joint reconstruction and fusion



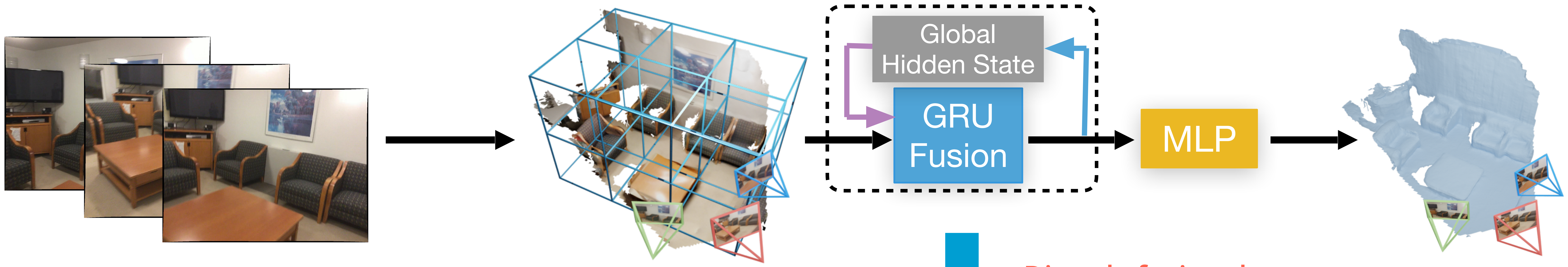
Input:
Posed Images in a fragment

View-Independent
Volume

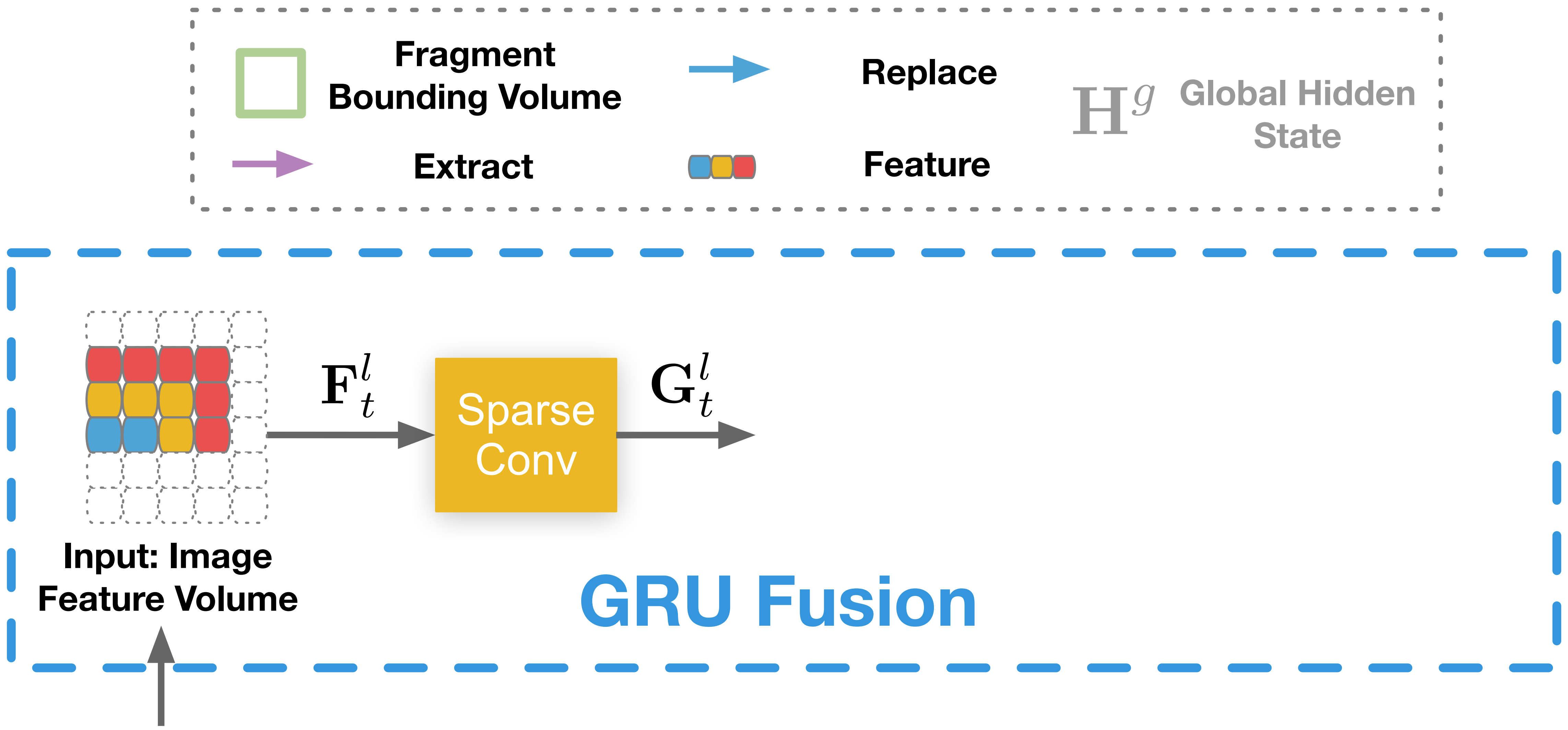
Fragment Geometry
(sparse TSDF volume)

NeuralRecon

Joint reconstruction and fusion

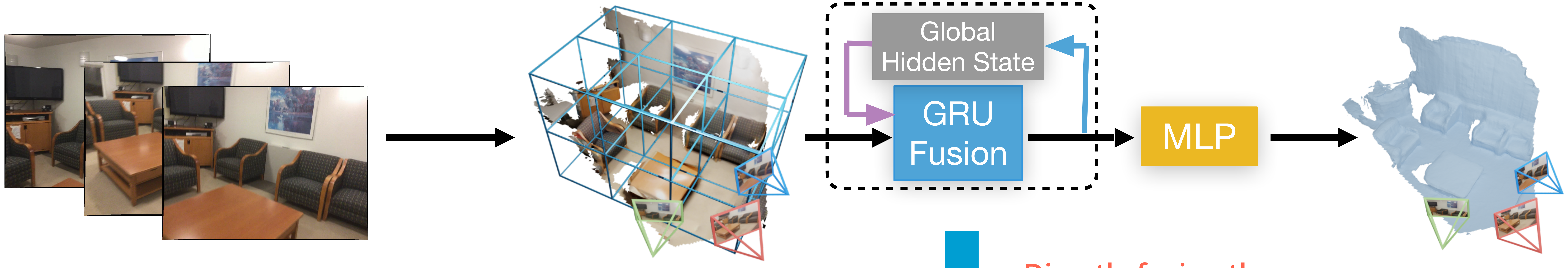


Directly fusing the features with GRU

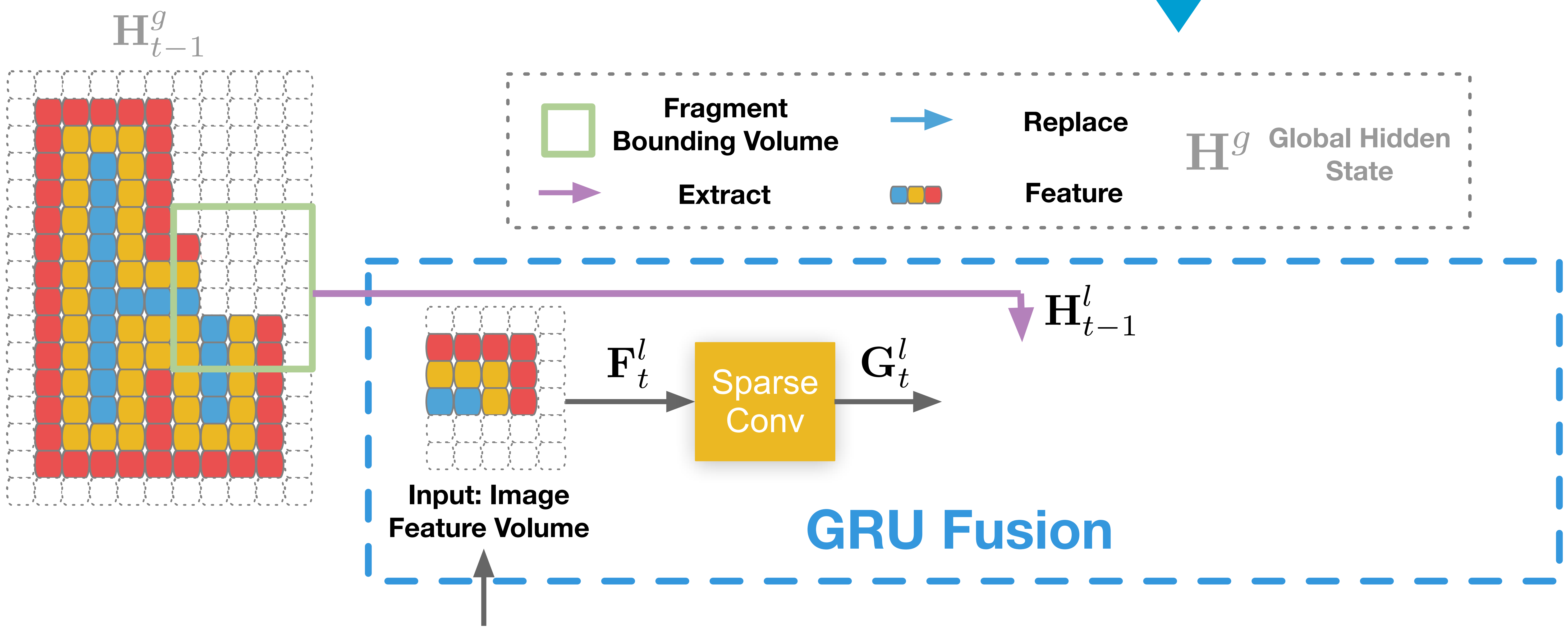


NeuralRecon

Joint reconstruction and fusion

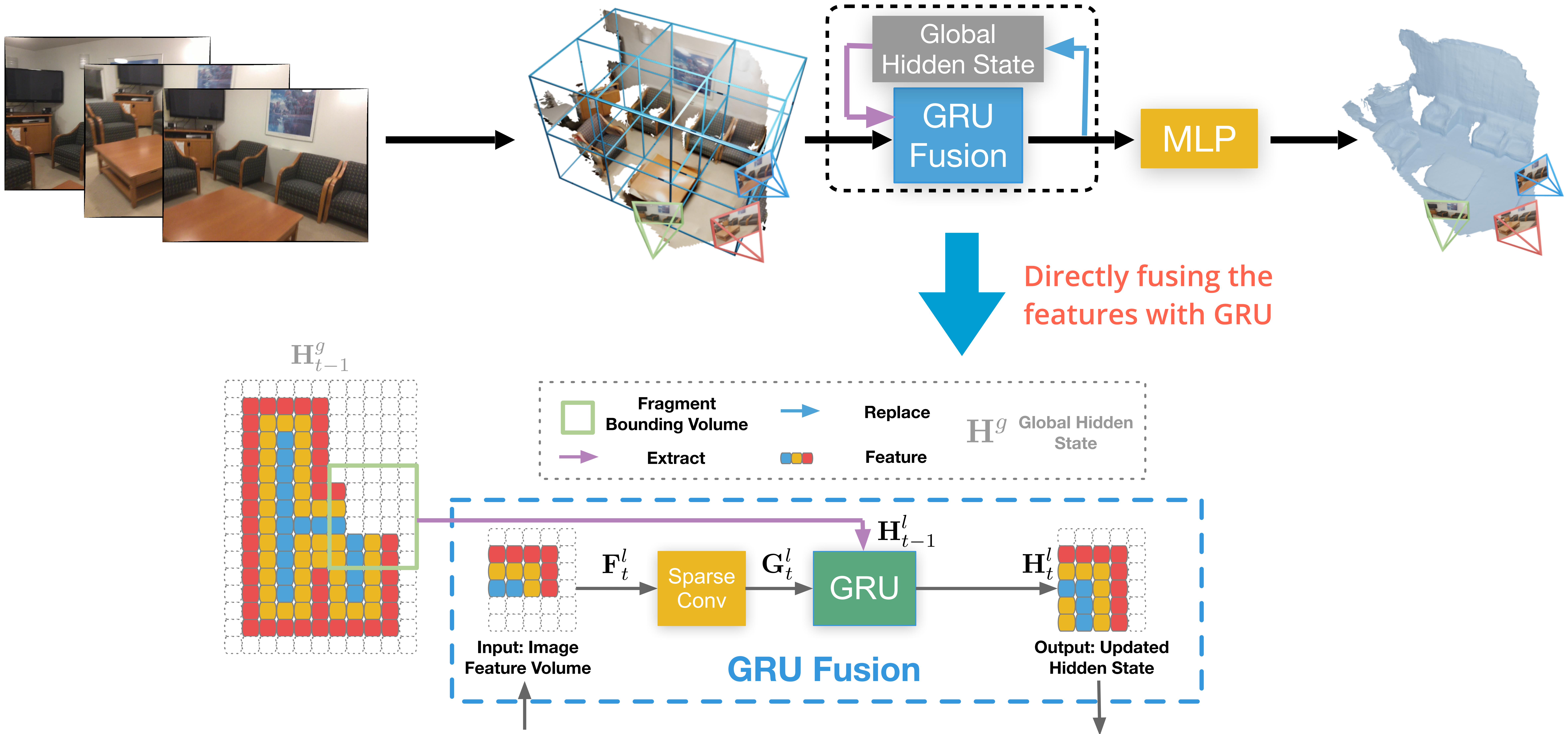


Directly fusing the features with GRU



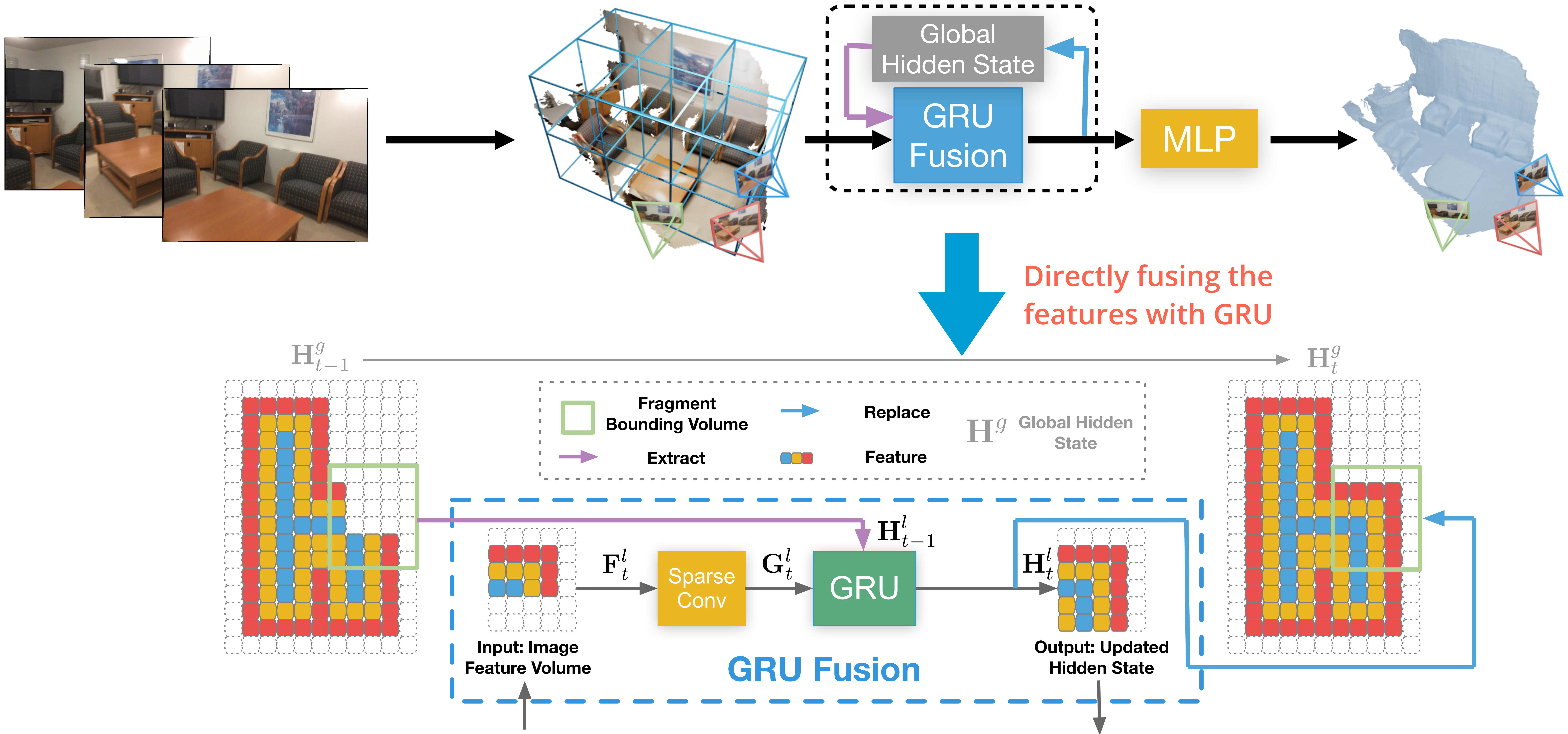
NeuralRecon

Joint reconstruction and fusion



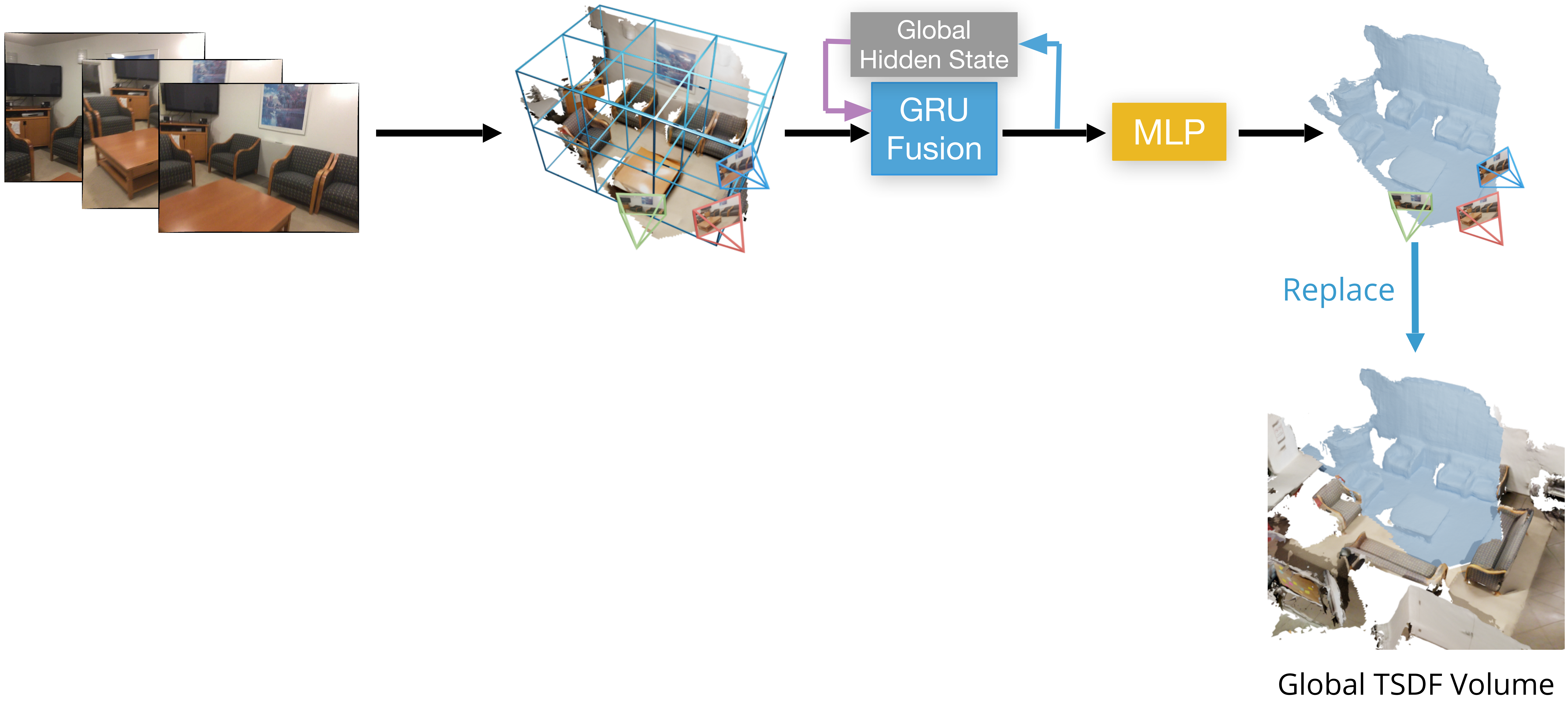
NeuralRecon

Joint reconstruction and fusion



NeuralRecon

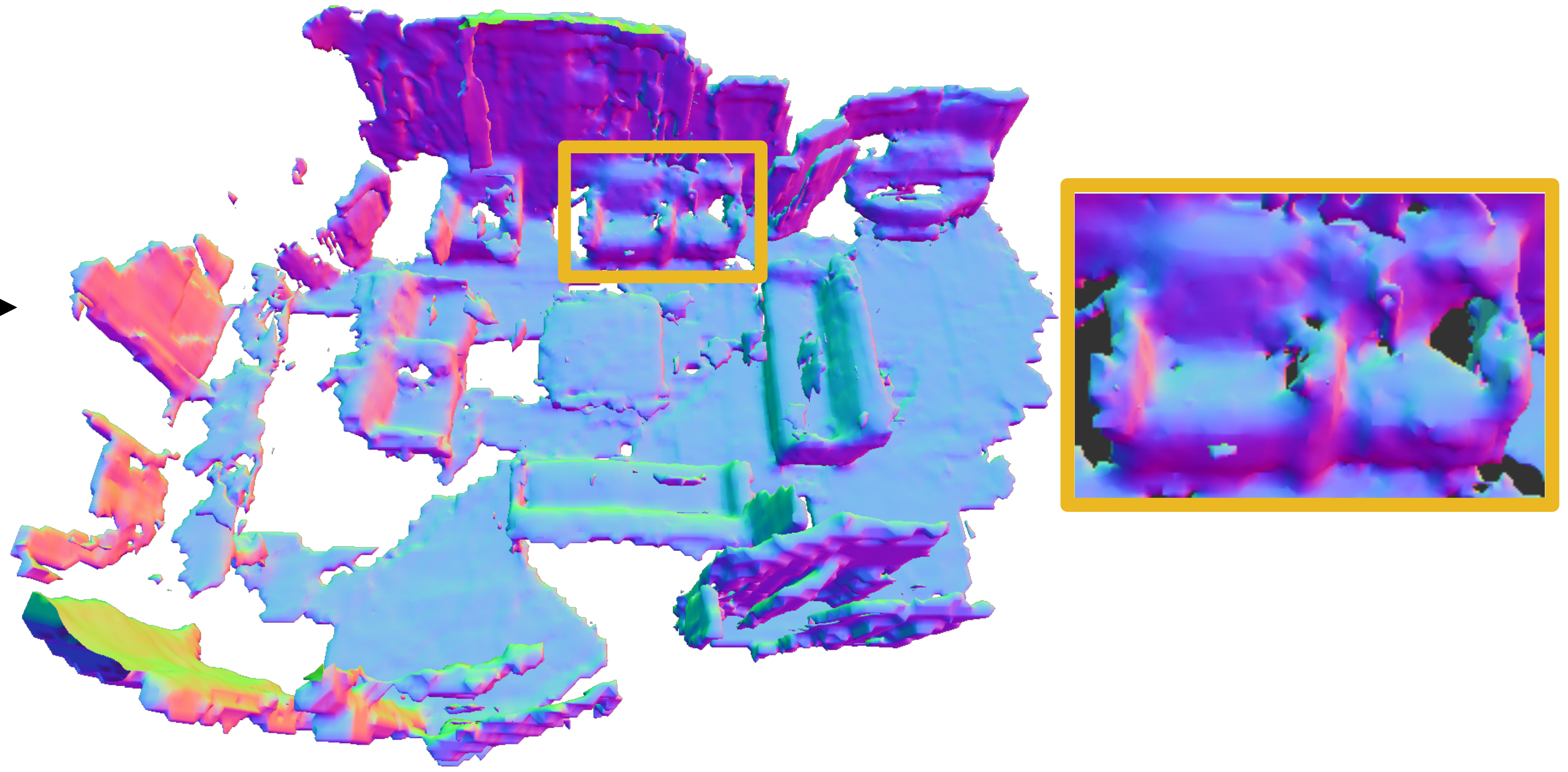
Joint reconstruction and fusion



NeuralRecon

Joint reconstruction and fusion

Fragment Reconstruction → TSDF Fusion →



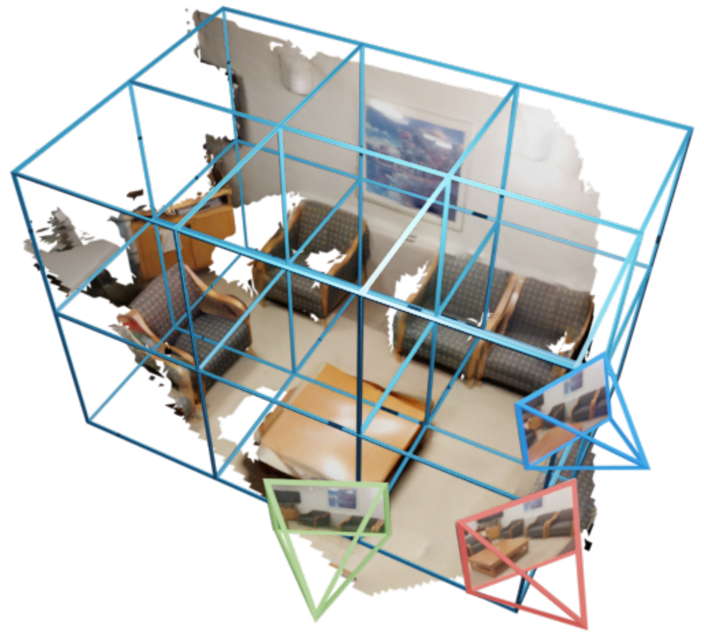
VS

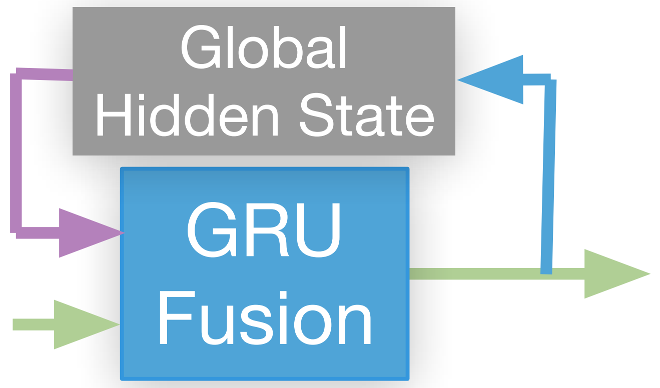
😃 Globally coherent

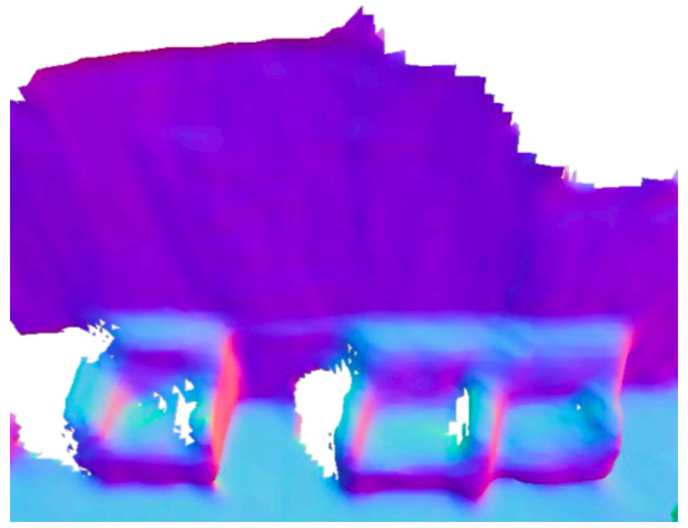
Joint Reconstruction and Fusion →

NeuralRecon

Conclusion







View-Independent Volume

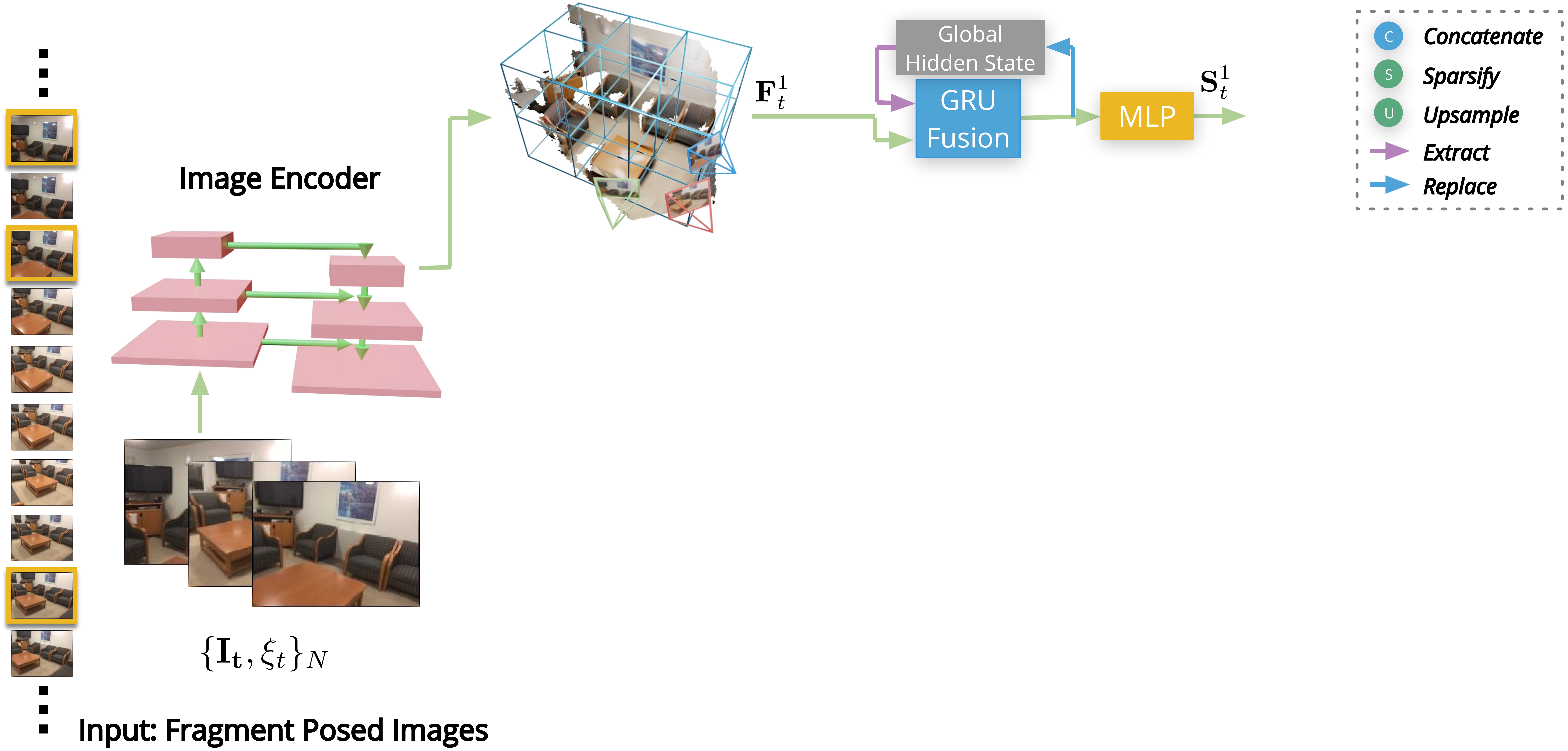
Joint TSDF Reconstruction and Fusion

Real-time

Coherent

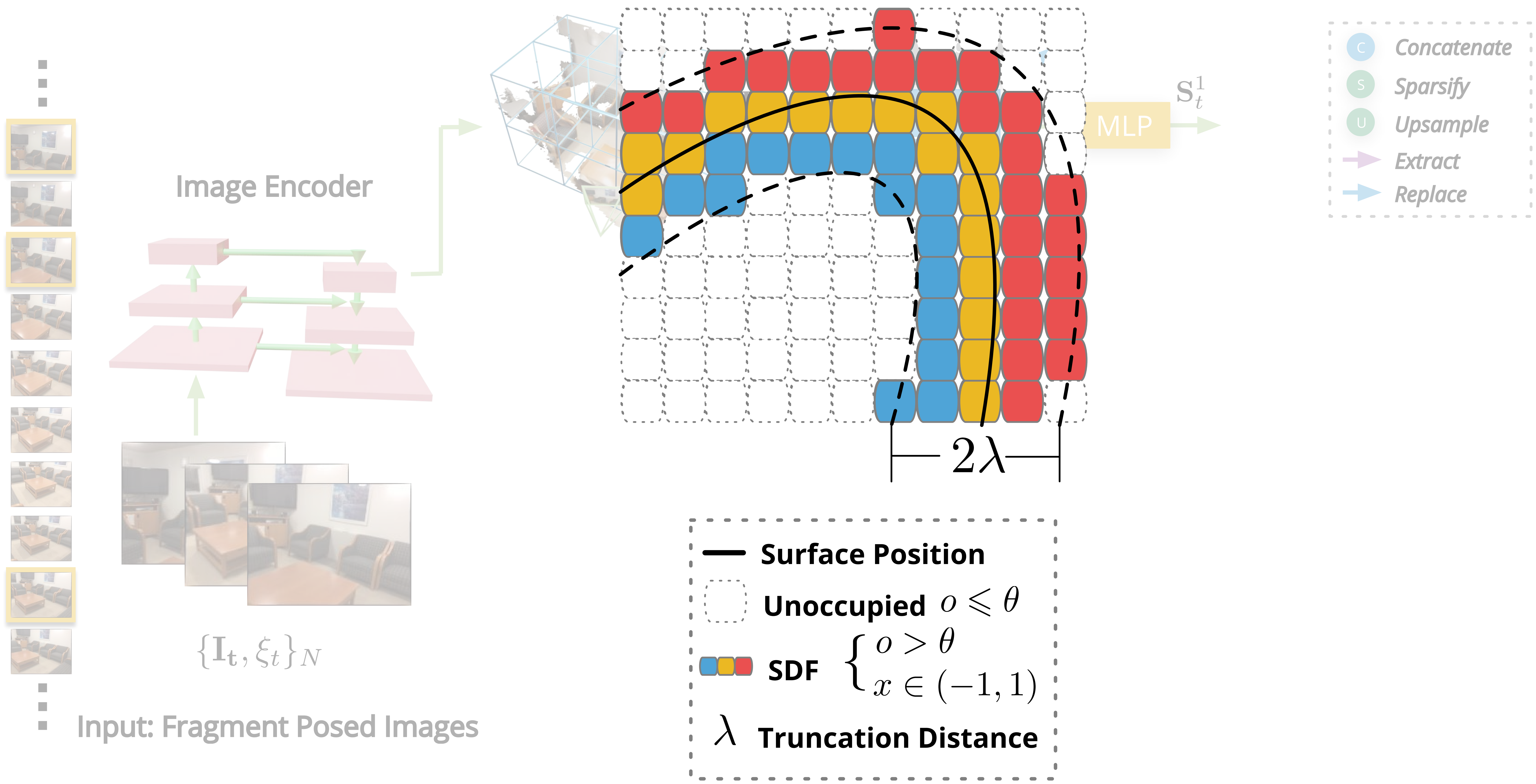
NeuralRecon

Coarse-to-fine architecture



NeuralRecon

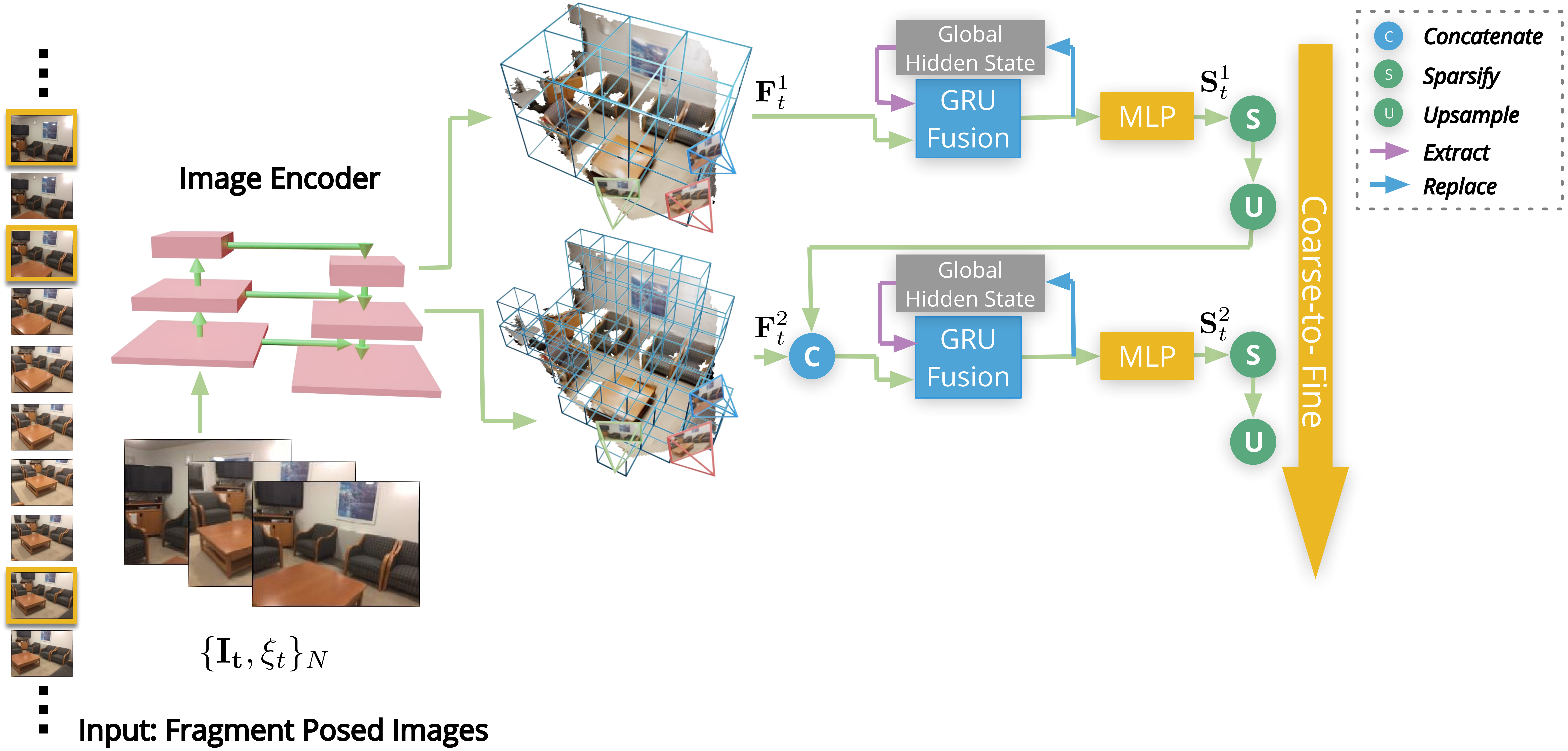
Coarse-to-fine architecture



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

NeuralRecon

Coarse-to-fine architecture

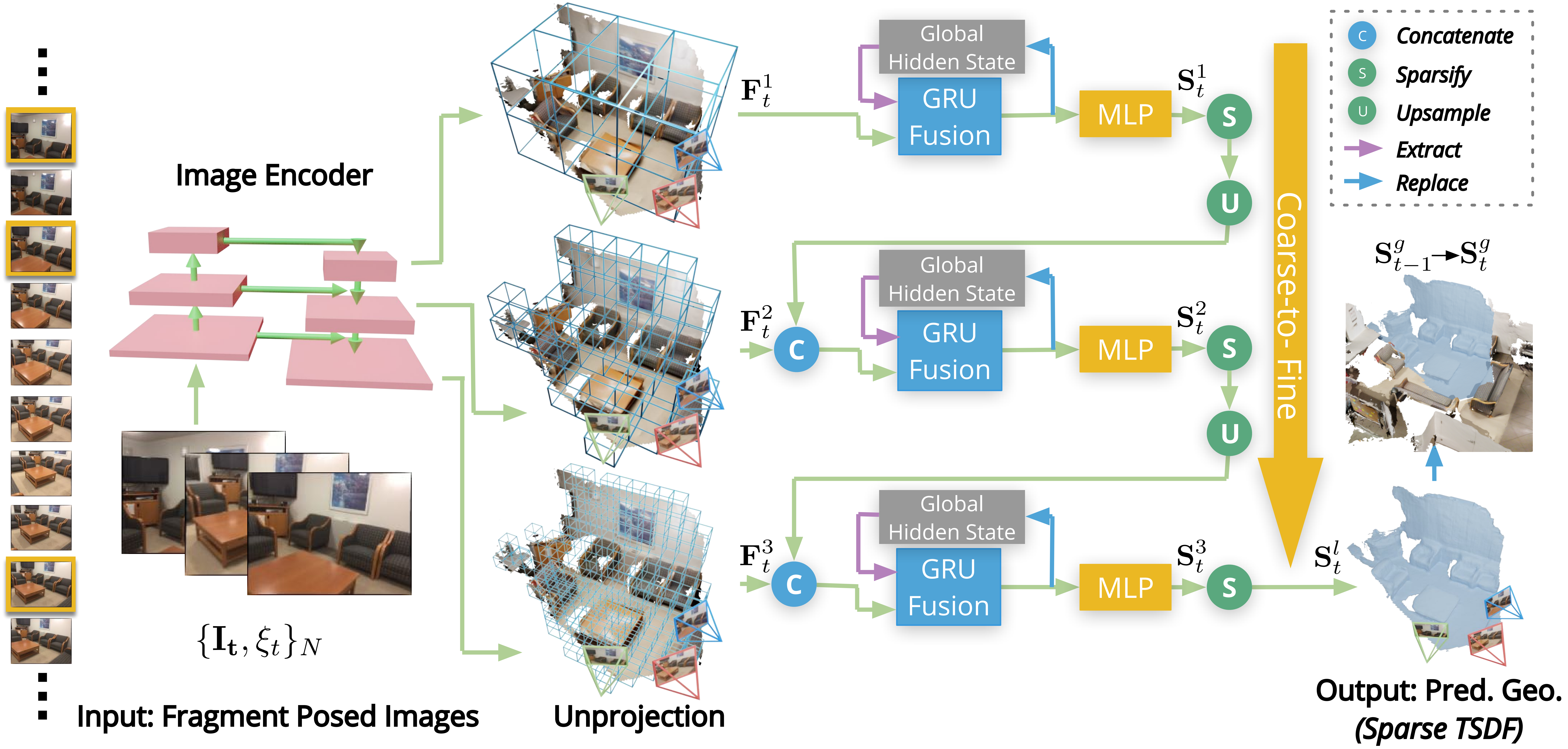


Output of *MLP* : **Occupancy Score** and **TSDF**

S Filter by Occupancy Score > 0

NeuralRecon

Coarse-to-fine architecture

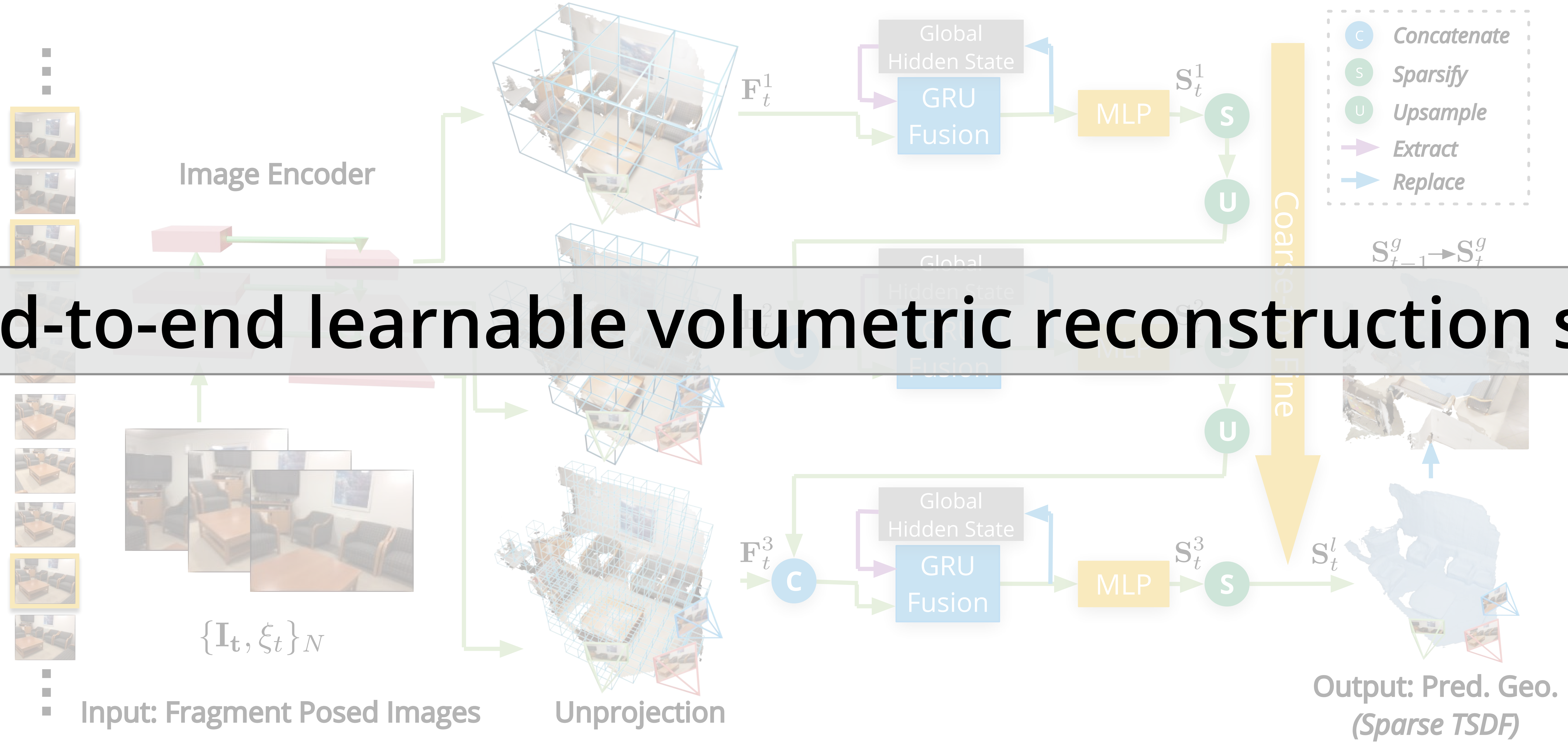


Output of *MLP* : **Occupancy Score** and **TSDF**

S Filter by Occupancy Score > 0

NeuralRecon

Coarse-to-fine architecture



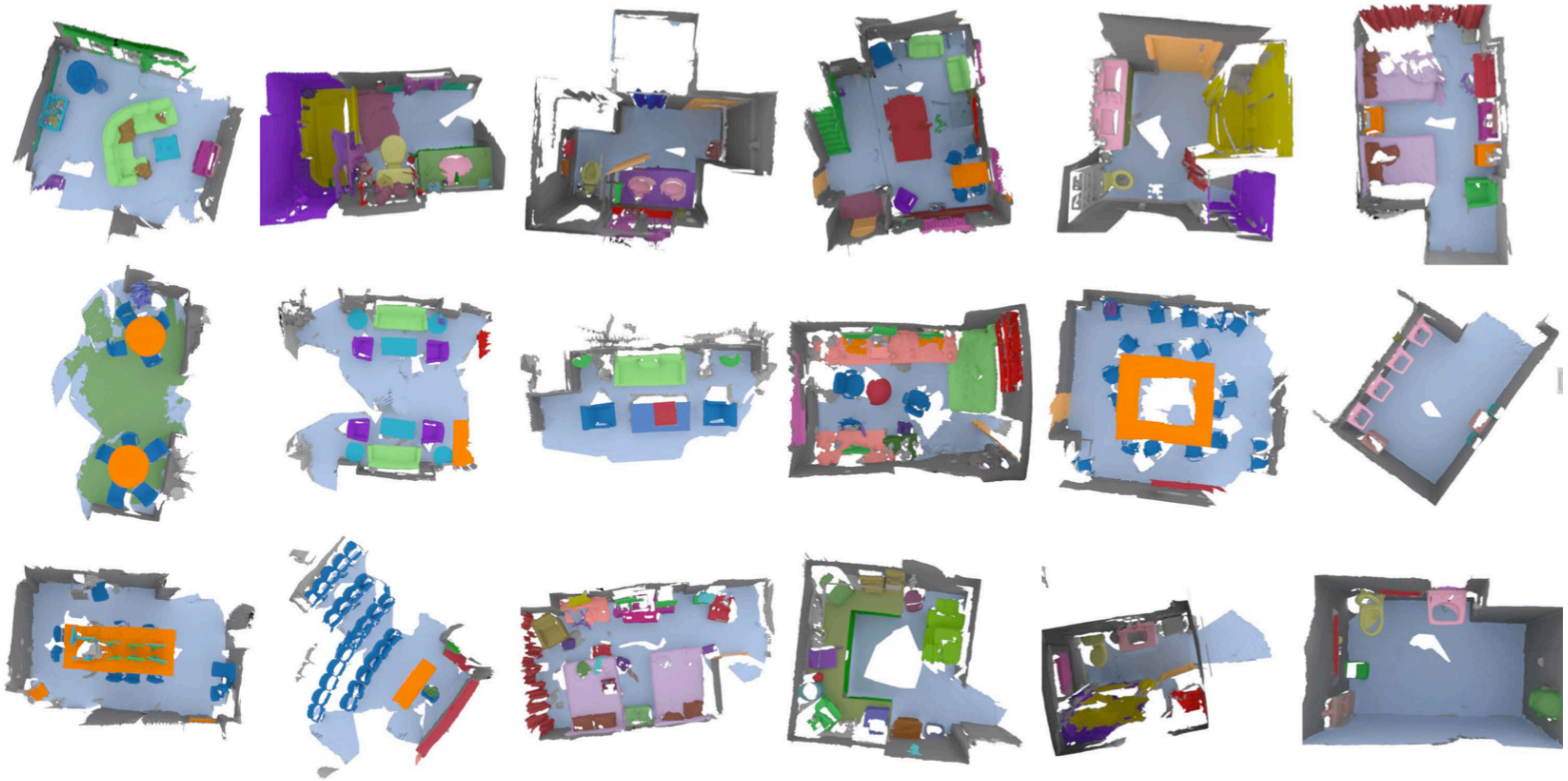
End-to-end learnable volumetric reconstruction system

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

S Filter by Occupancy Score > 0

NeuralRecon

Training



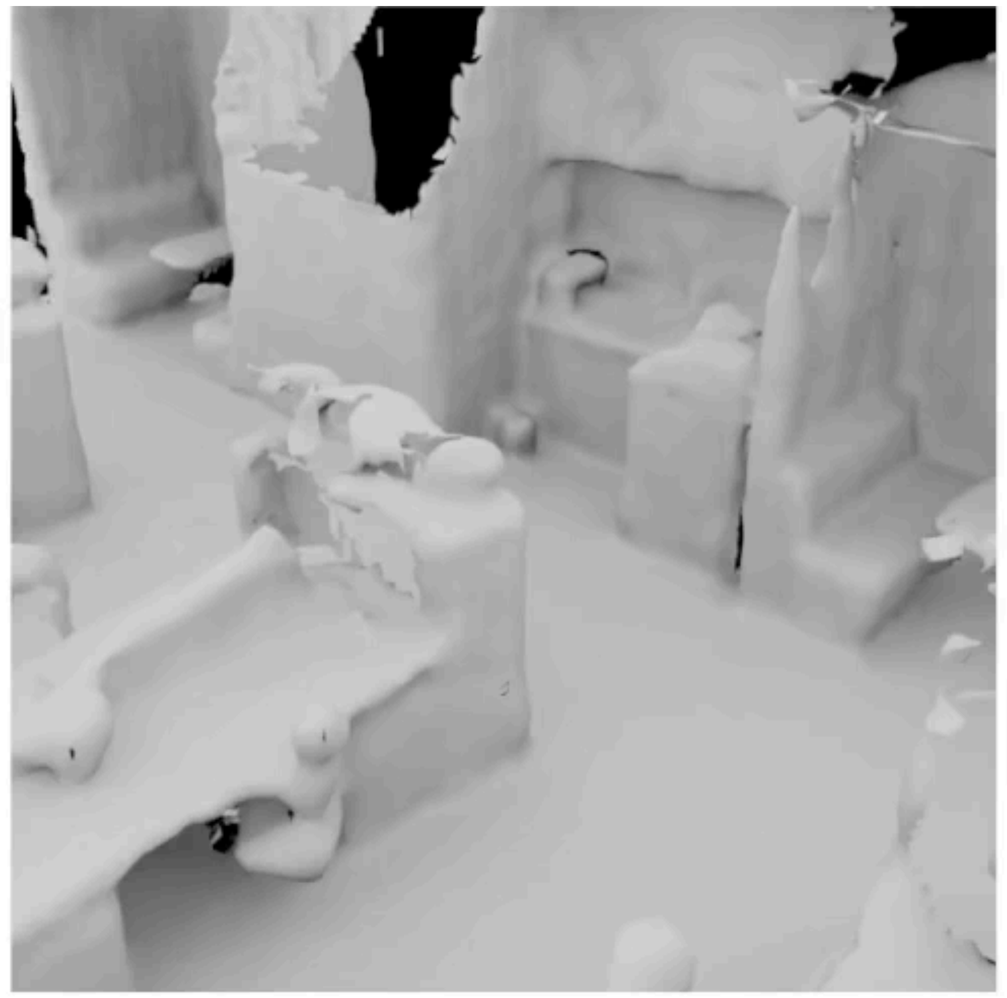
ScanNet dataset

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

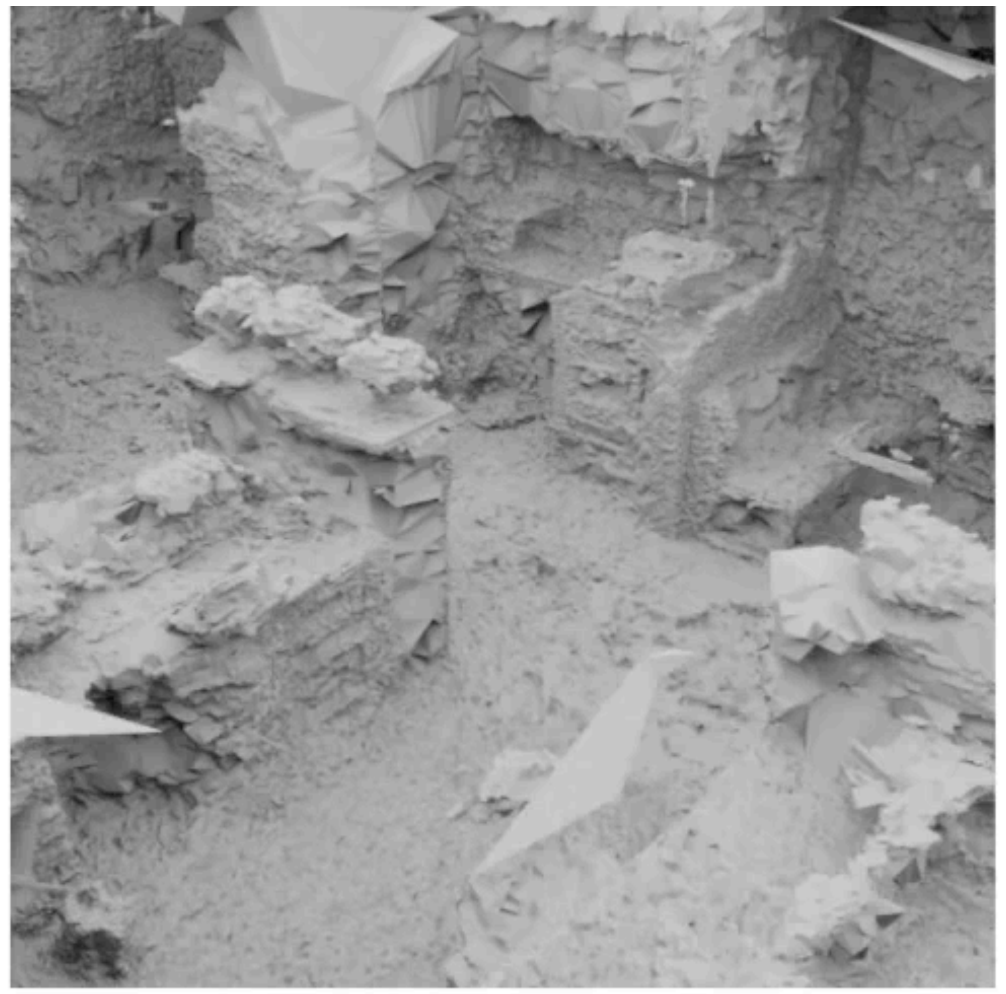
Binary cross-entropy (BCE) and L1 loss are used for training

Experiments

Qualitative results: office 1

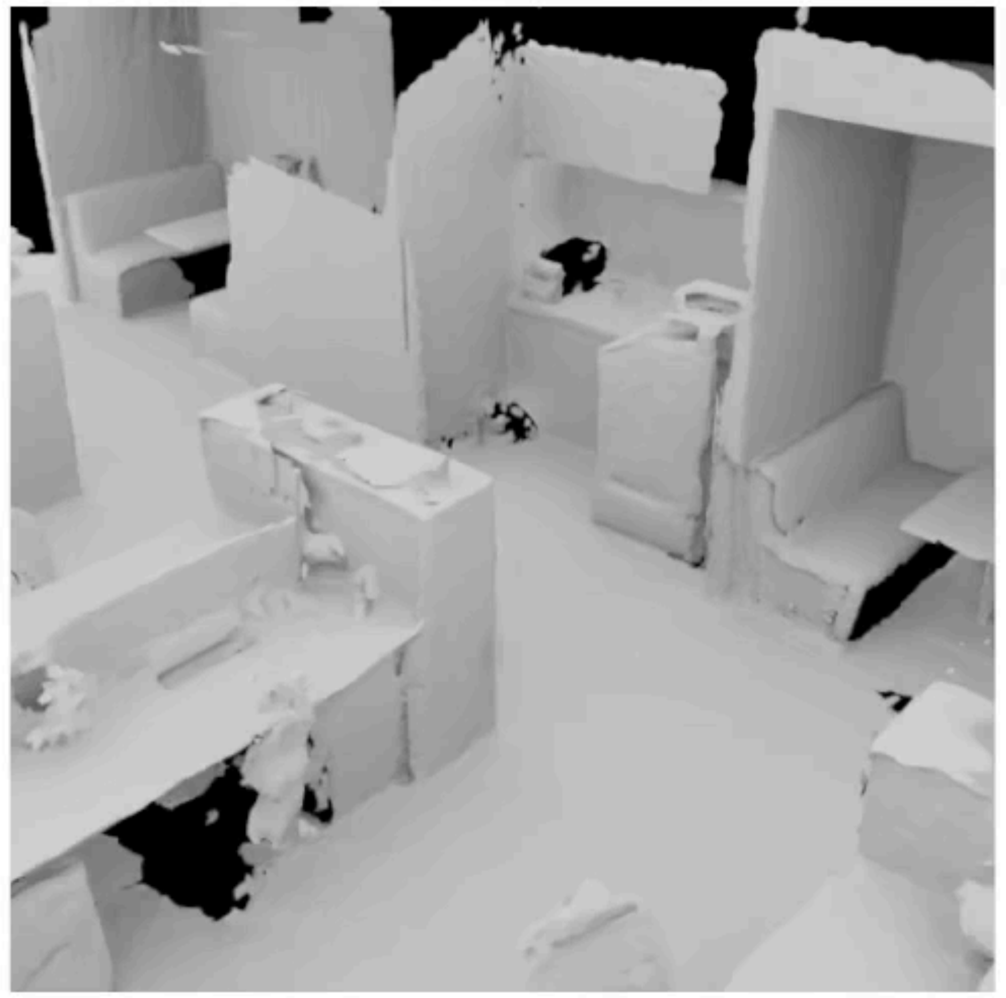


Ours (30ms)

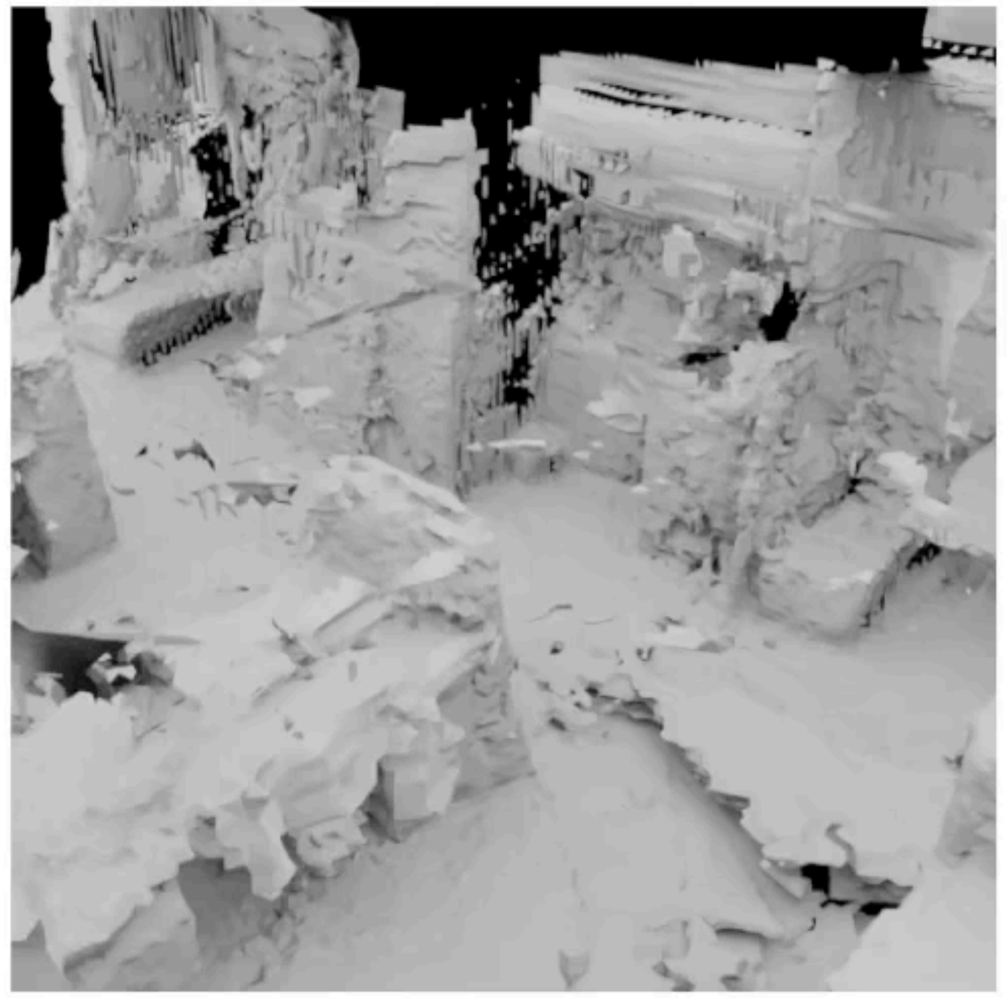


COLMAP (2076ms)

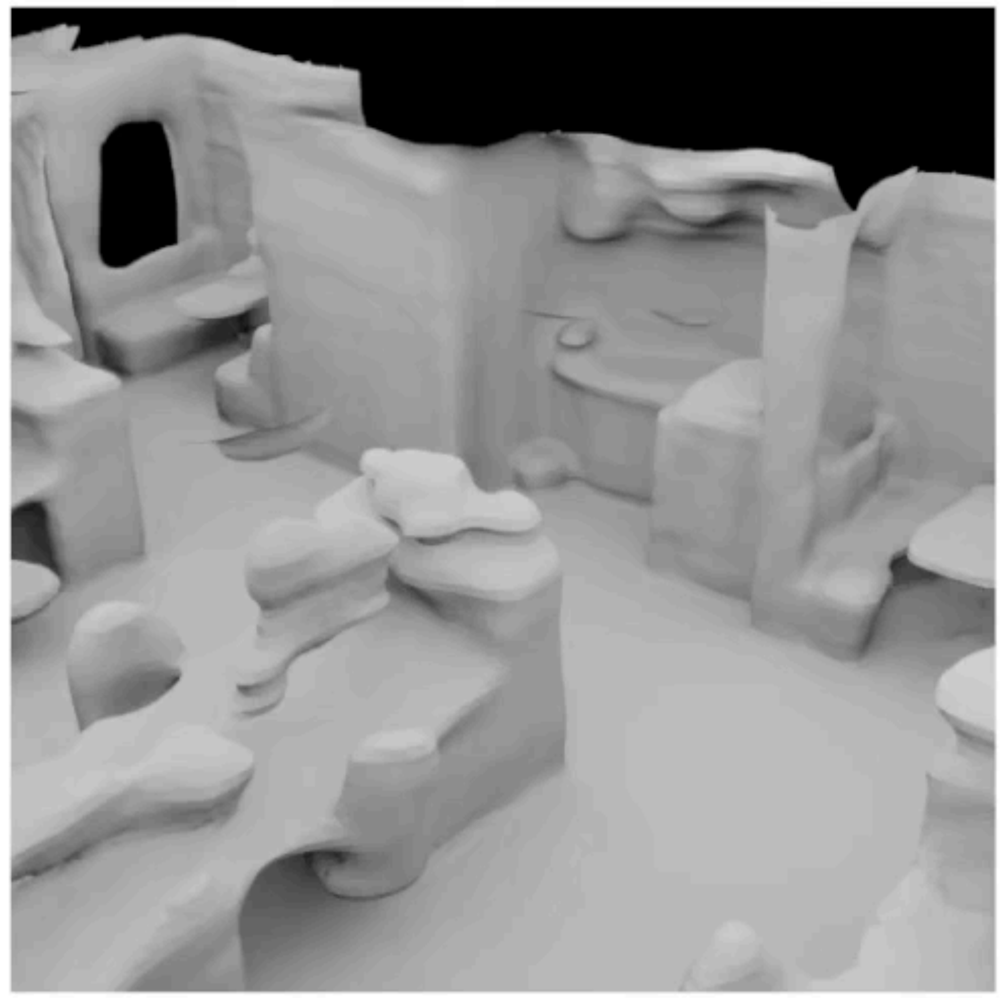
DeepV2D (347ms)



Ground Truth



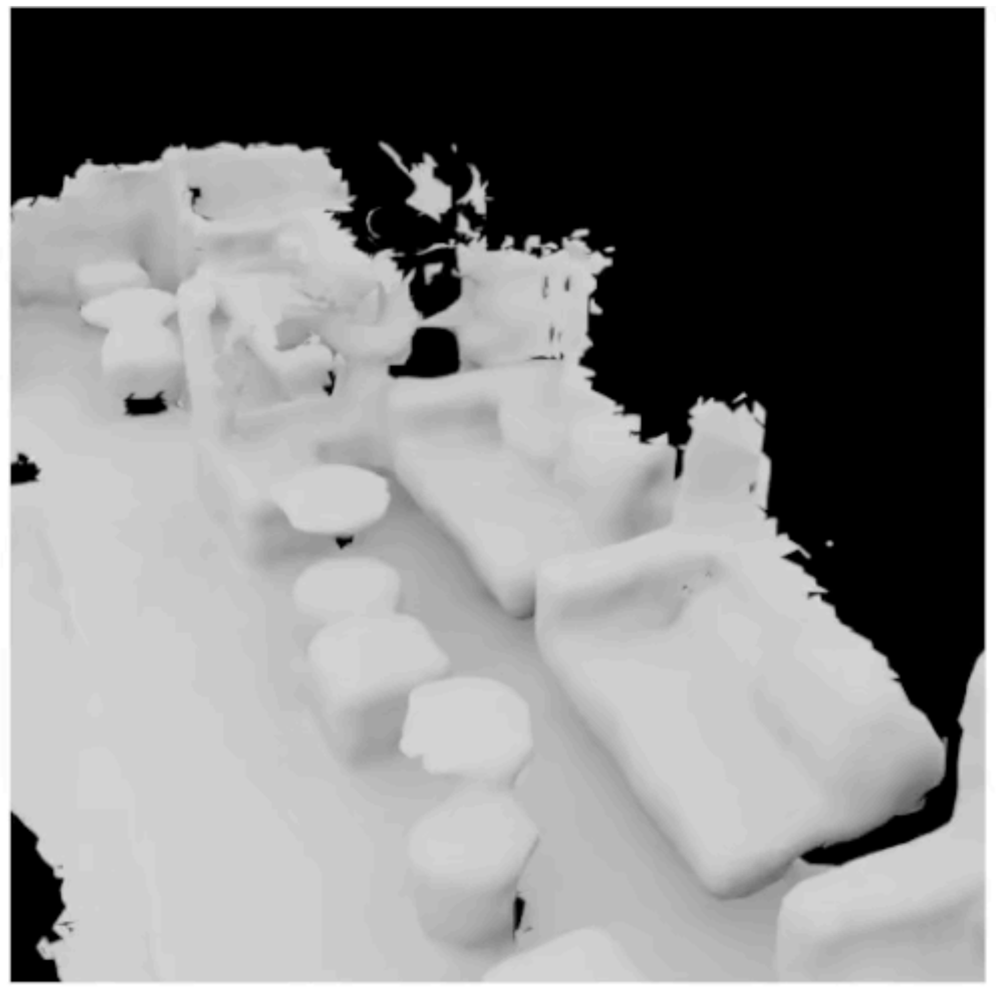
CNMNet (80ms)



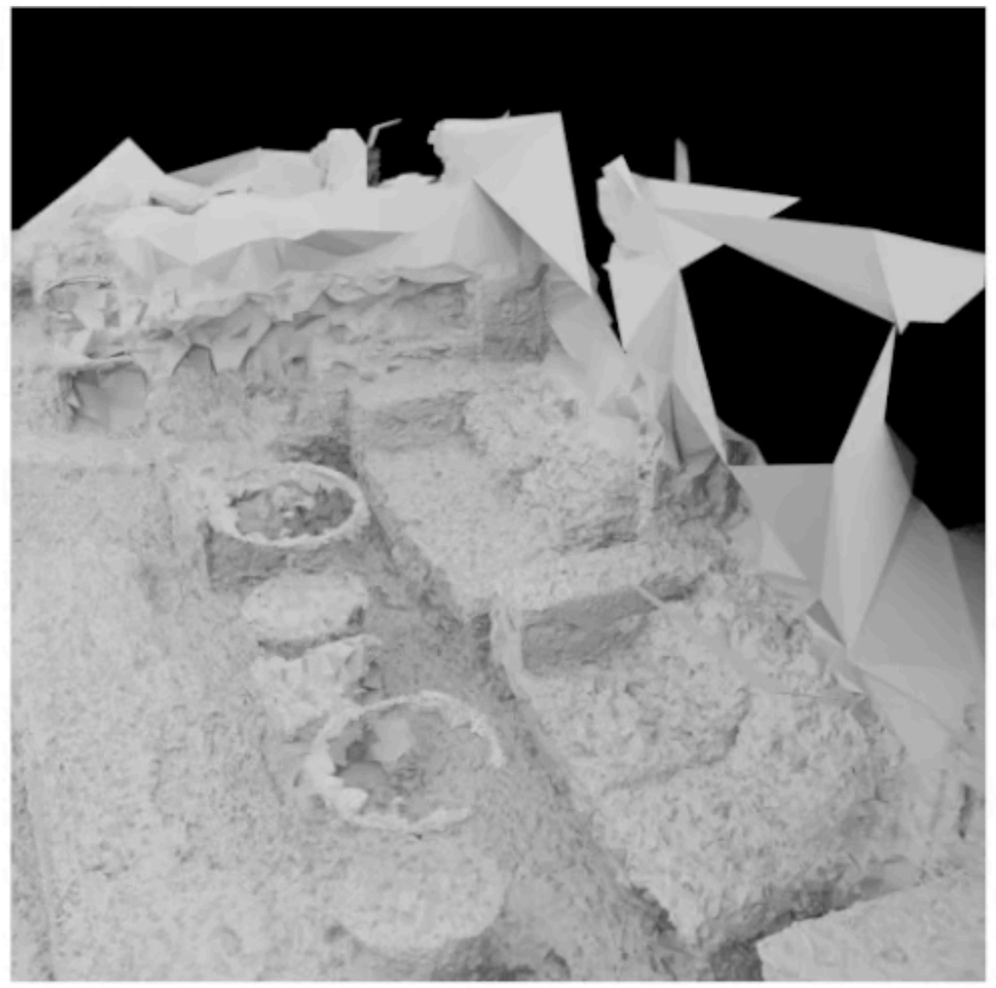
Atlas (292ms)

Experiments

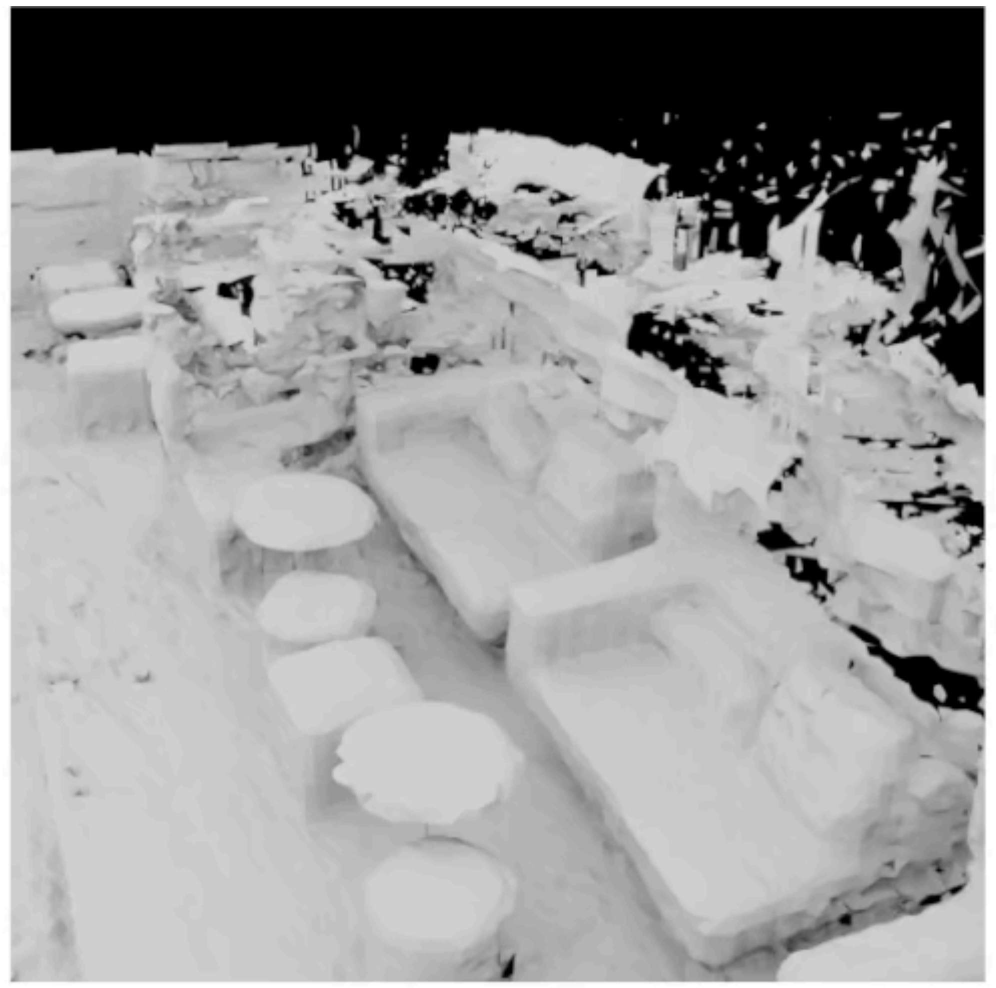
Qualitative results: office 2



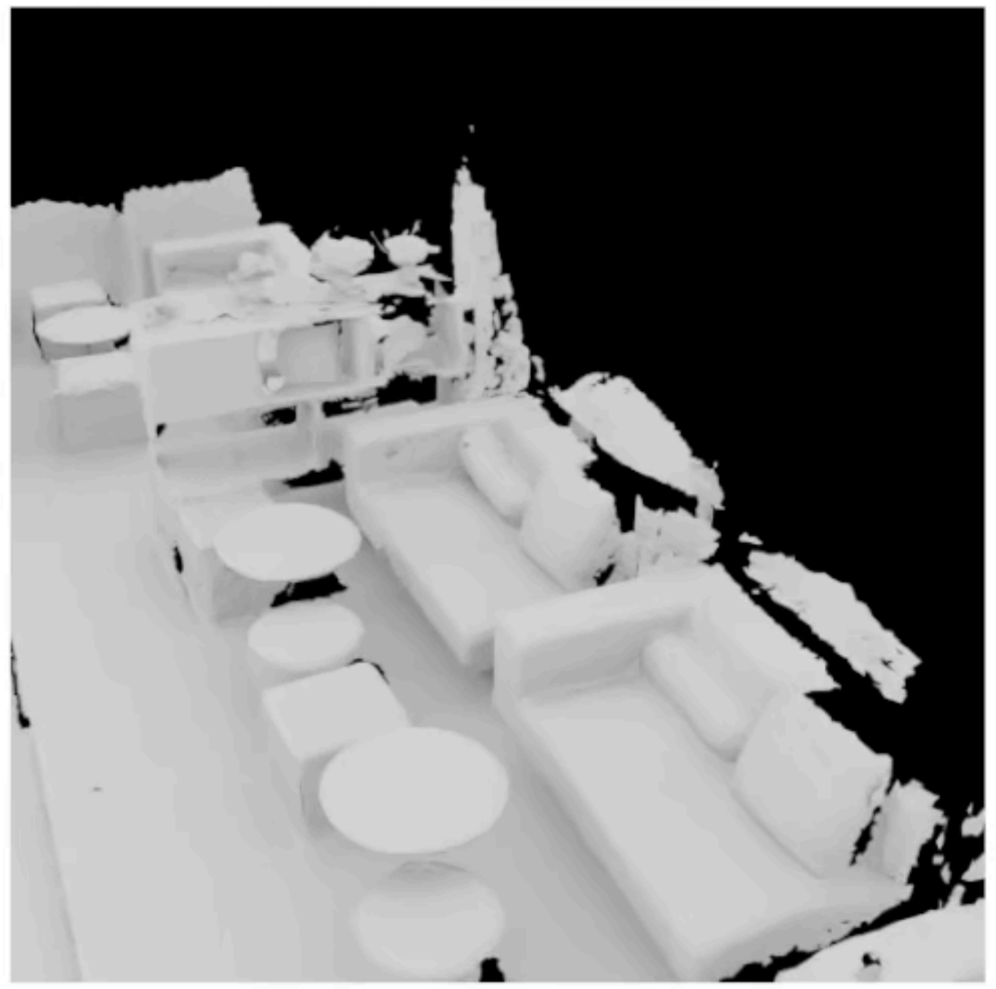
Ours (30ms)



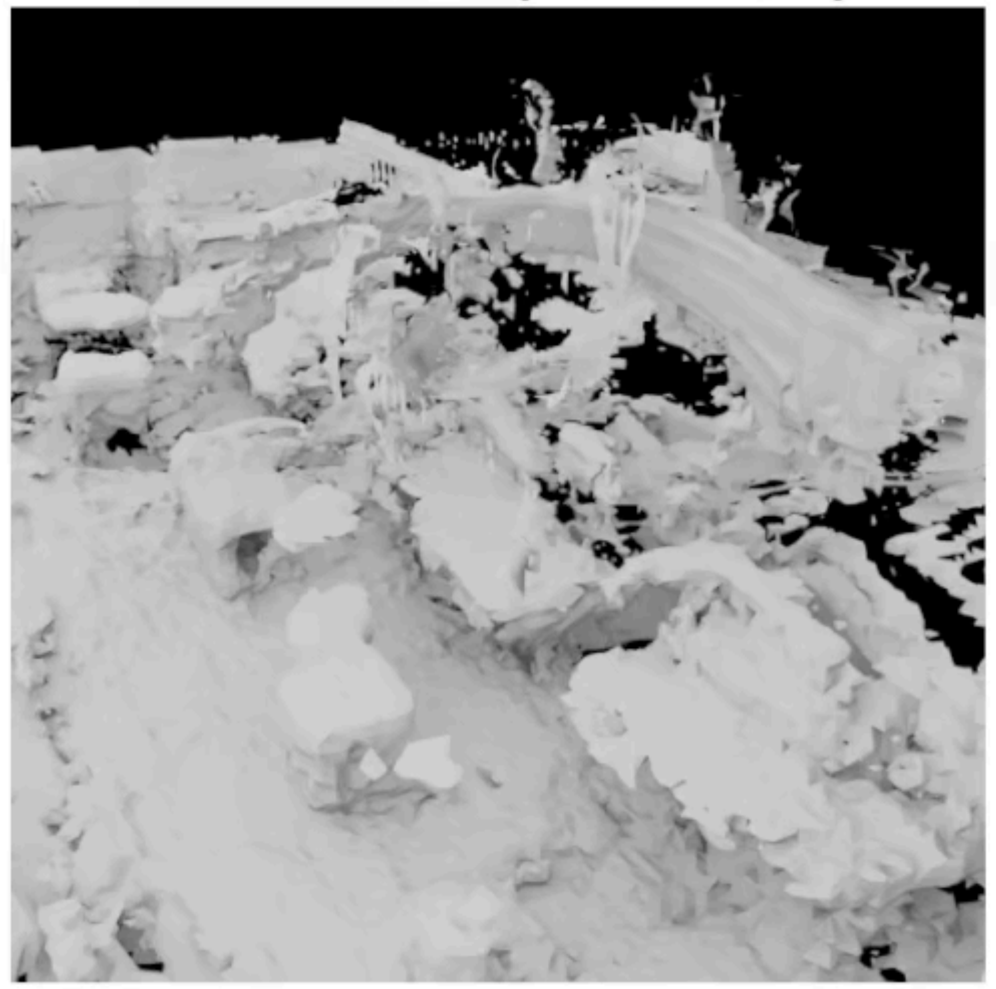
COLMAP (2076ms)



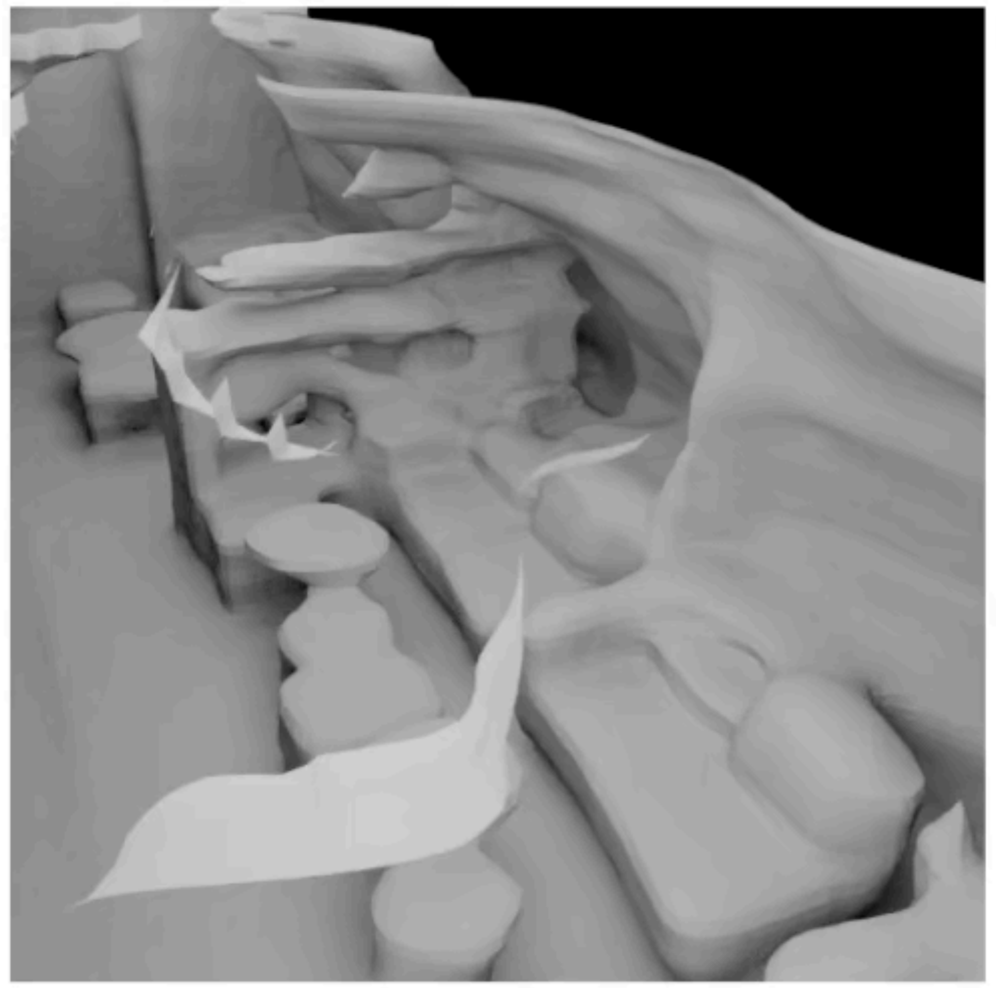
DeepV2D (347ms)



Ground Truth



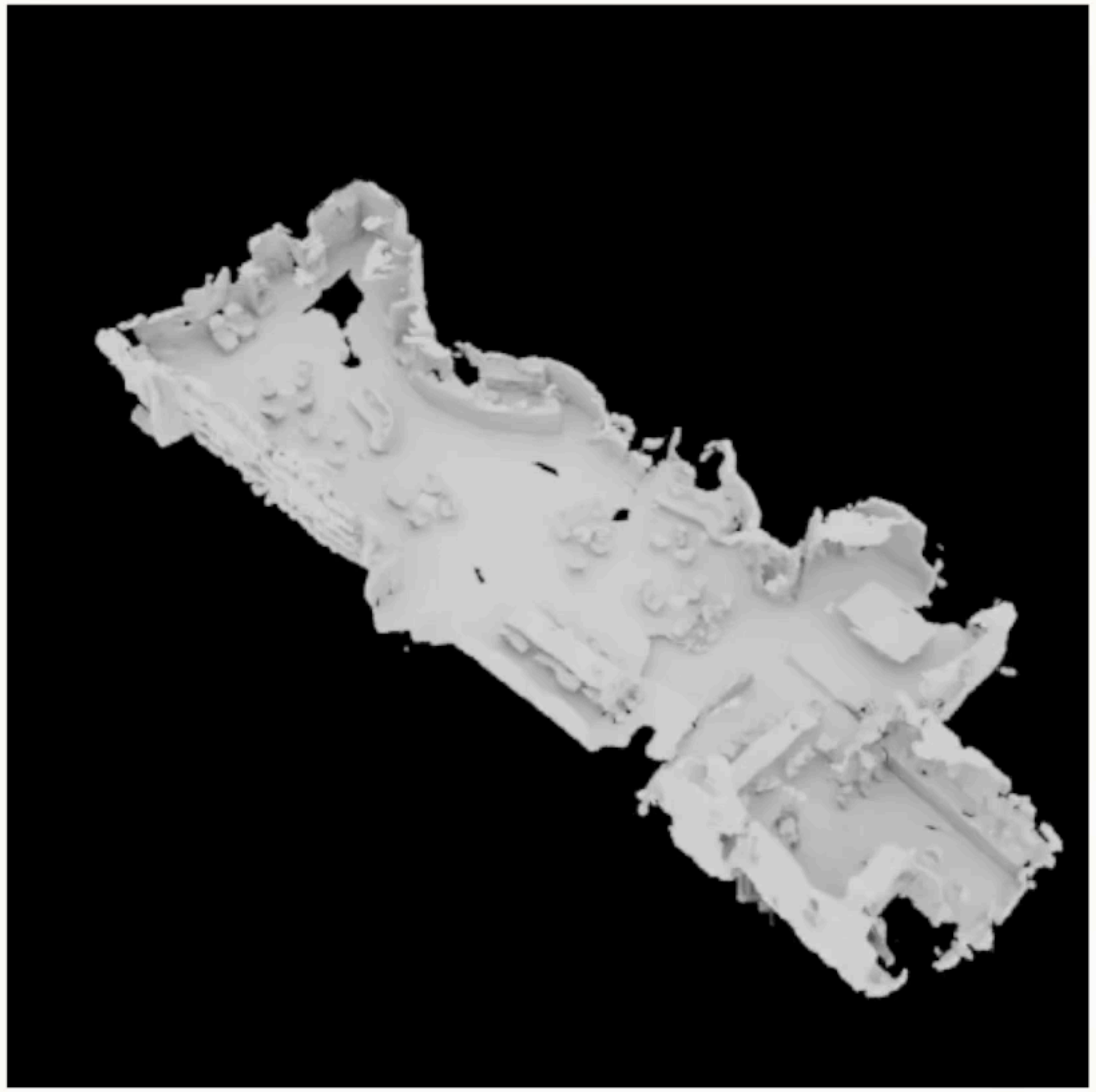
CNMNet (80ms)



Atlas (292ms)

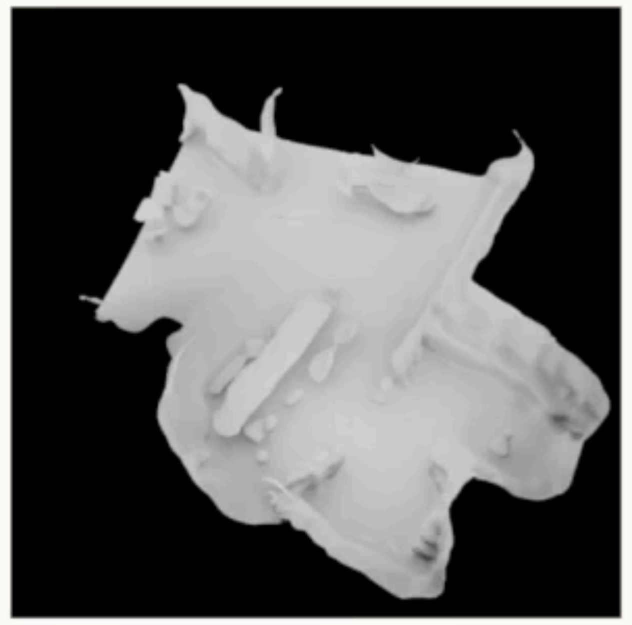
Experiments

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



Ours

Max GPU Memory: 3.29GB



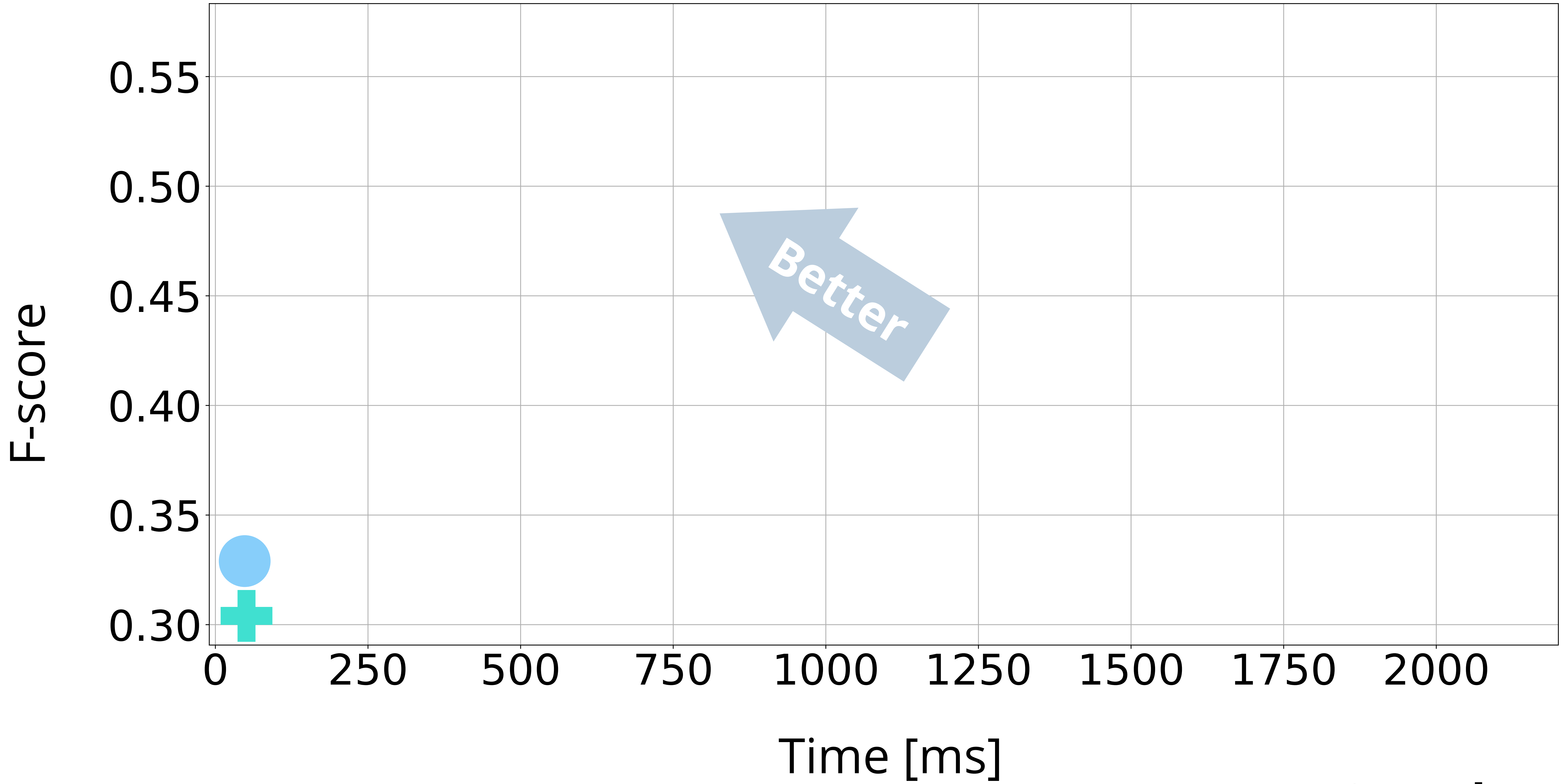
Atlas

Max GPU Memory: >24GB (OOM)

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

Experiments

Quantitive results



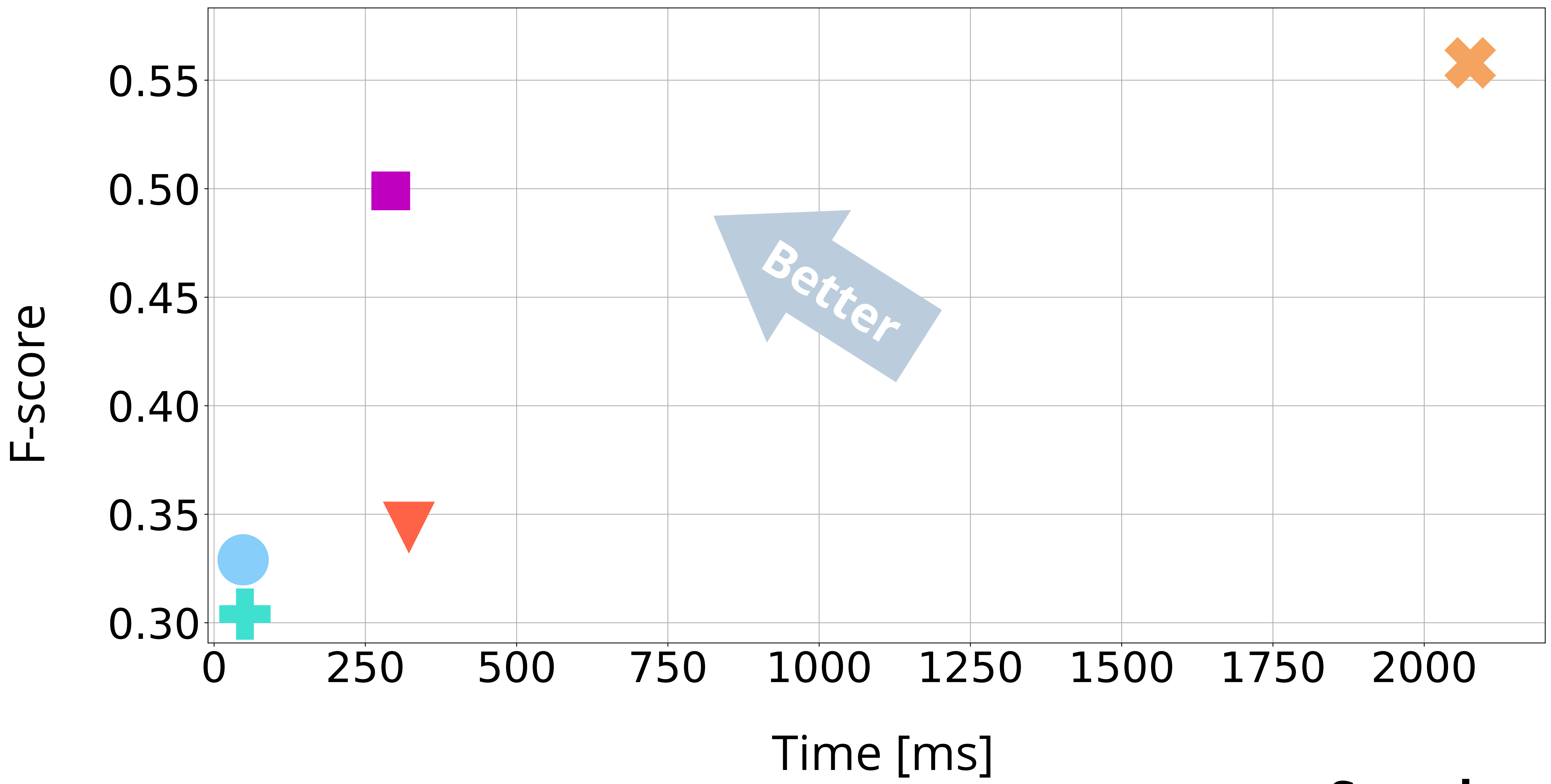
Real-time methods: MVDepthNet GPMVS

Speed 😃

Accuracy 🙁

Experiments

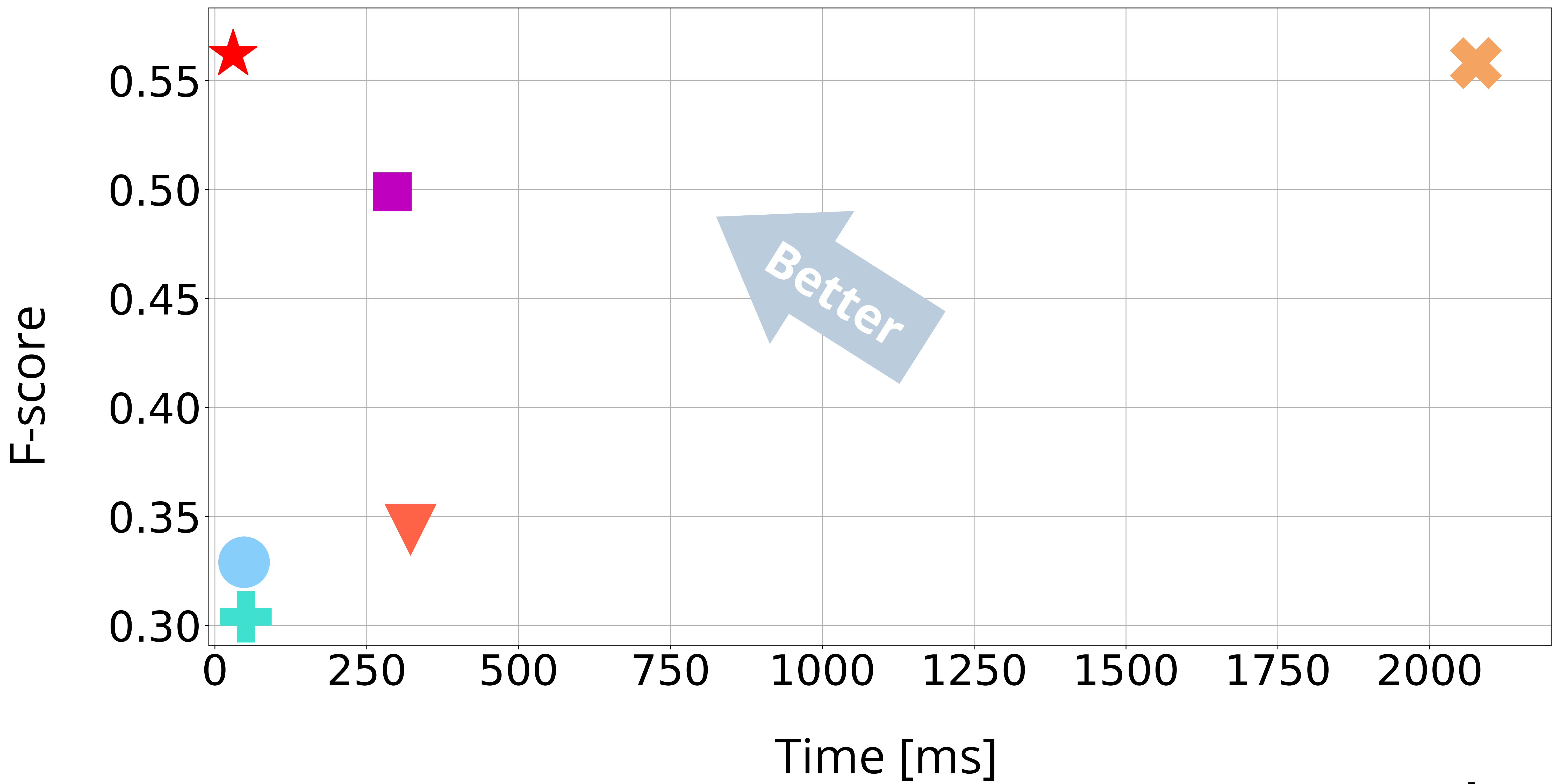
Quantitive results



				Speed	Accuracy
Real-time methods:	MVDepthNet	GPMVS		😃	🙁
Multiple View Stereo methods:	DPSNet	COLMAP	Atlas	🙁	😃

Experiments

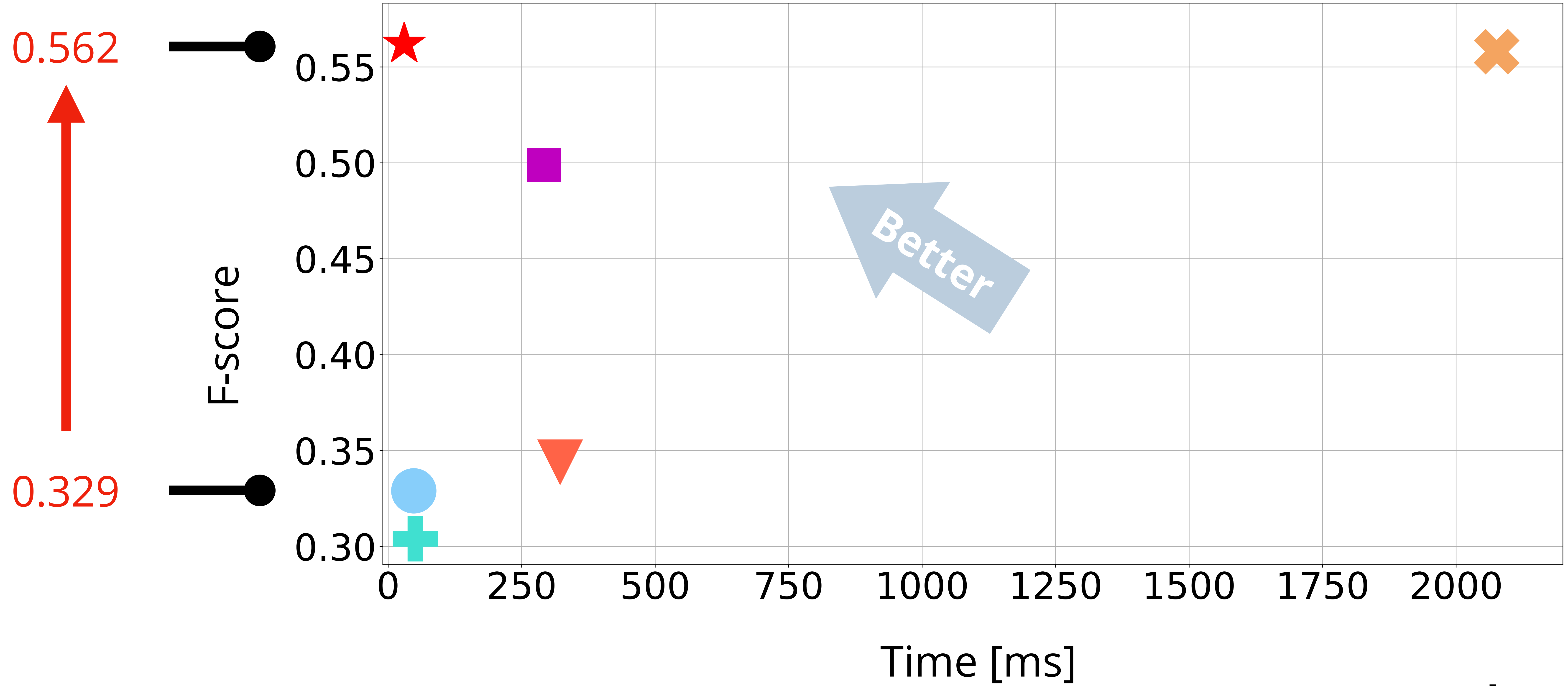
Quantitive results



				Speed	Accuracy
Real-time methods:	MVDepthNet	GPMVS		😃	🙁
Multiple View Stereo methods:	DPSNet	COLMAP	Atlas	🙁	😃
	Ours			😃	😃

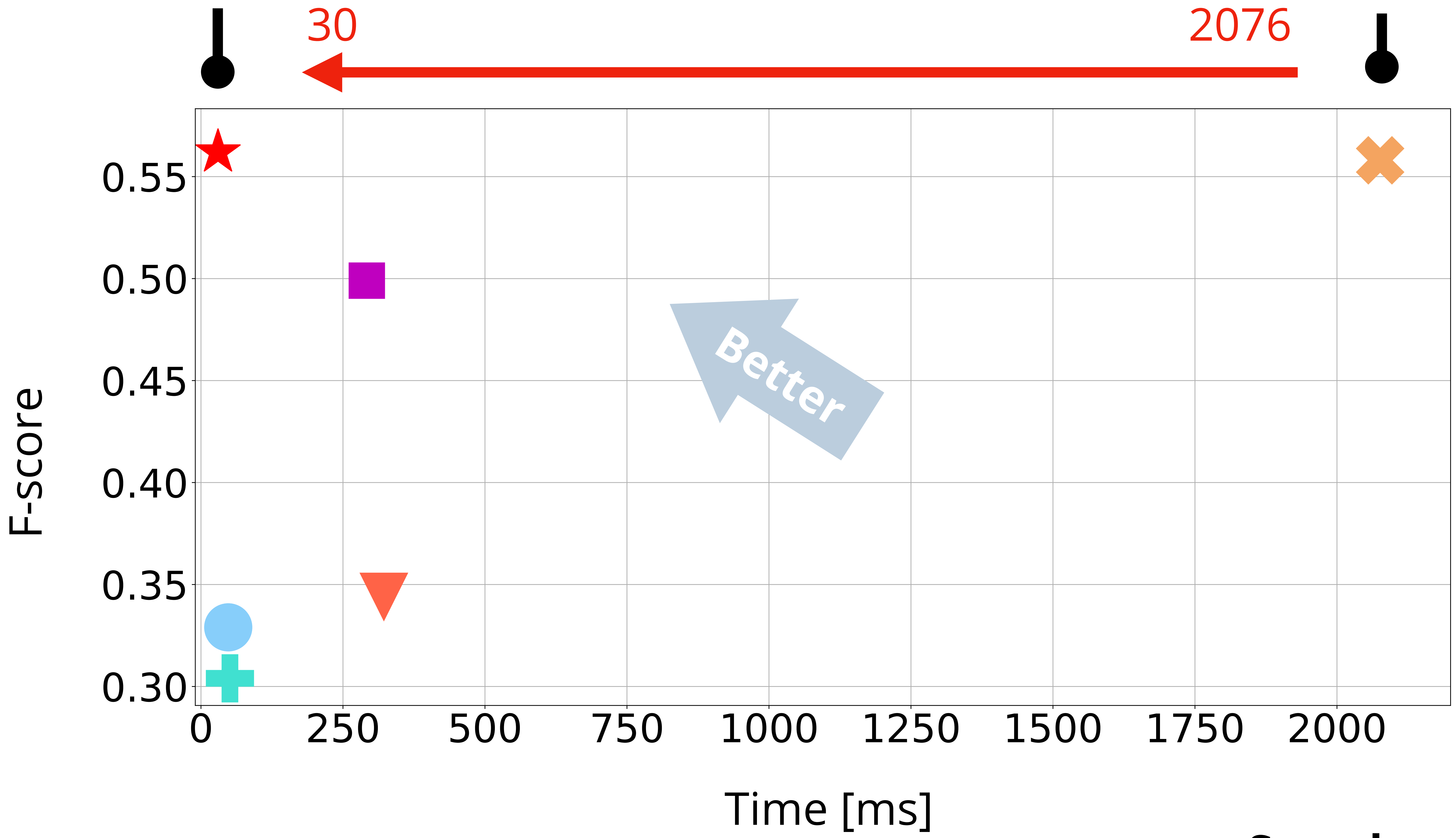
Experiments

Quantitive results



				Speed	Accuracy
Real-time methods:	MVDepthNet	GPMVS		😃	🙁
Multiple View Stereo methods:	DPSNet	COLMAP	Atlas	🙁	😃
	Ours			😃	😃

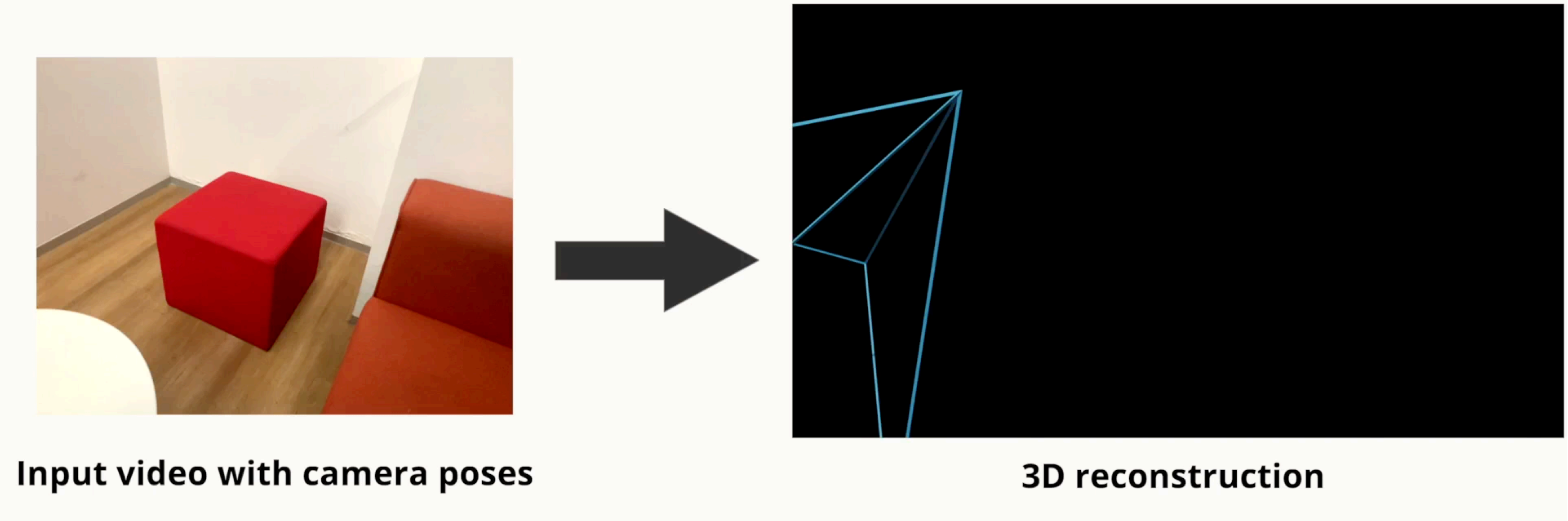
Experiments



				Speed	Accuracy
Real-time methods:	MVDepthNet	GPMVS		😃	🙁
Multiple View Stereo methods:	DPSNet	COLMAP	Atlas	🙁	😃
	Ours			😃	😃

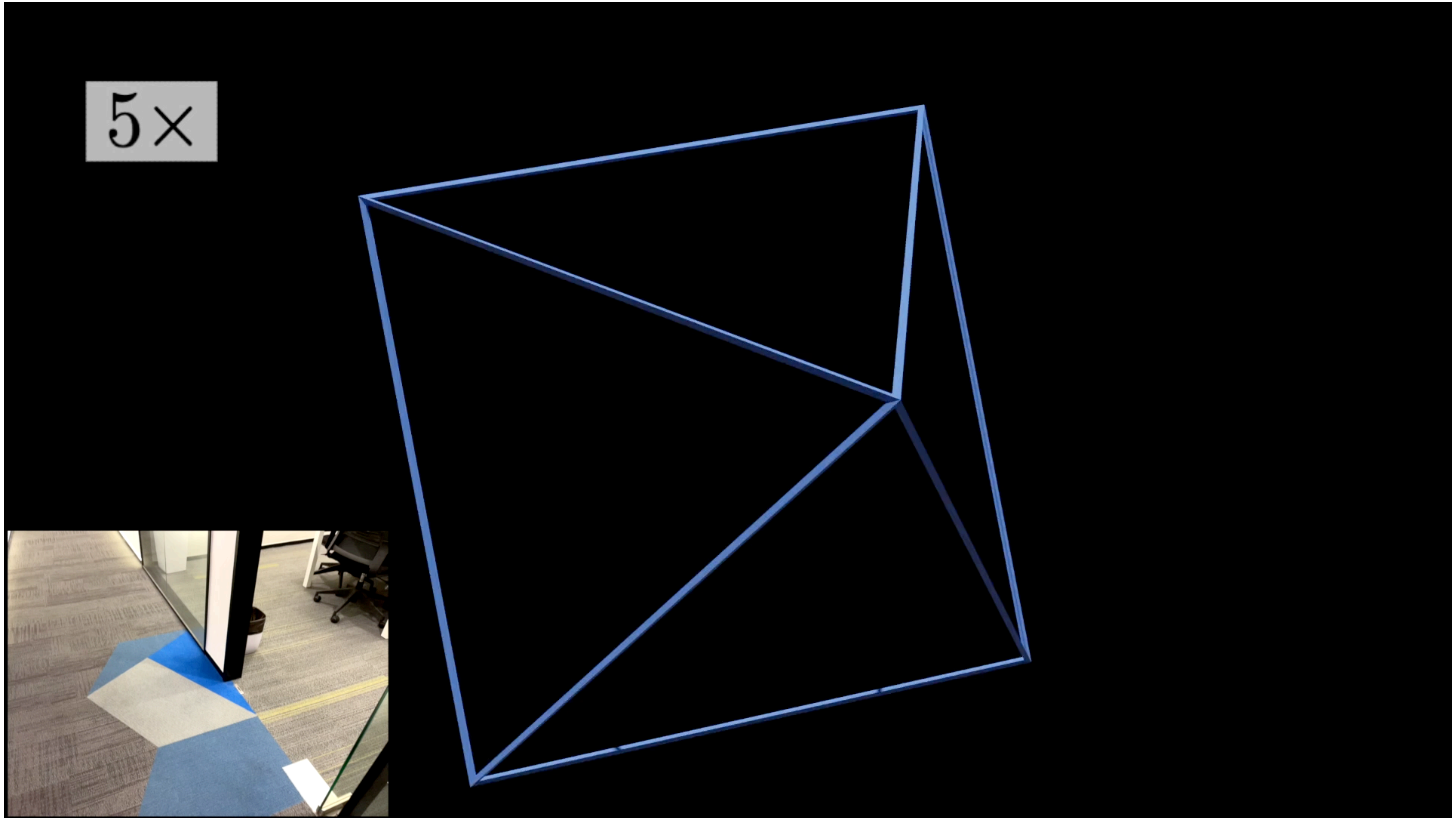
Demo

Indoor scene at our office



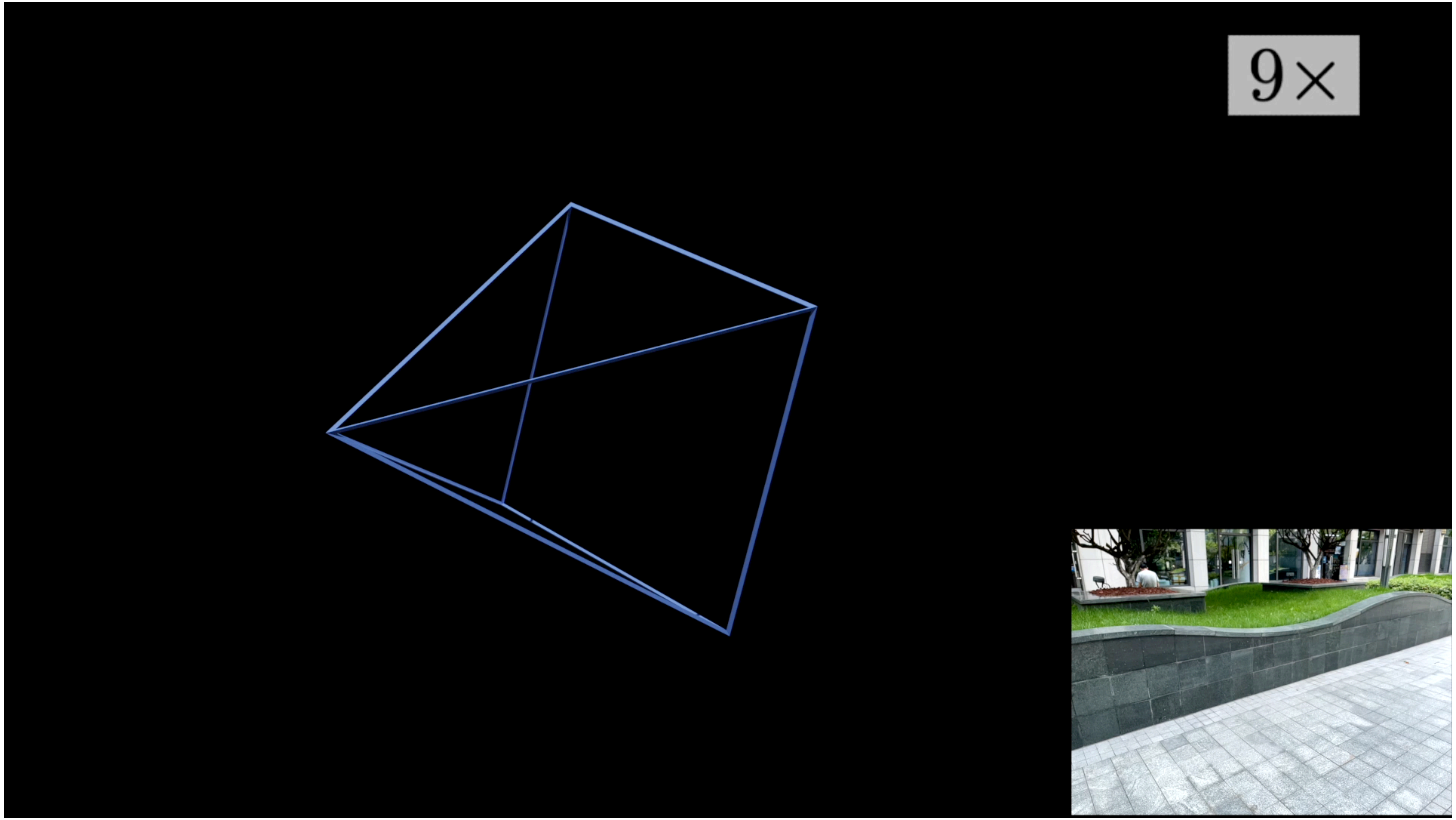
Demo

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/

Paper link: https://arxiv.org/pdf/2104.00681.pdf

Thanks for watching!
Q&A