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

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

CVPR 2021 (Oral)

Presenter: Yiming Xie

商汤
sensetime

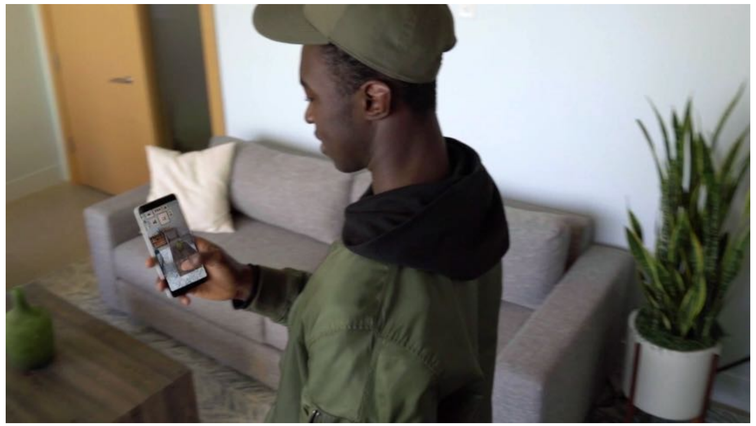
Background

Task: 3D Reconstruction from monocular video

Input video with camera poses

3D reconstruction

Background



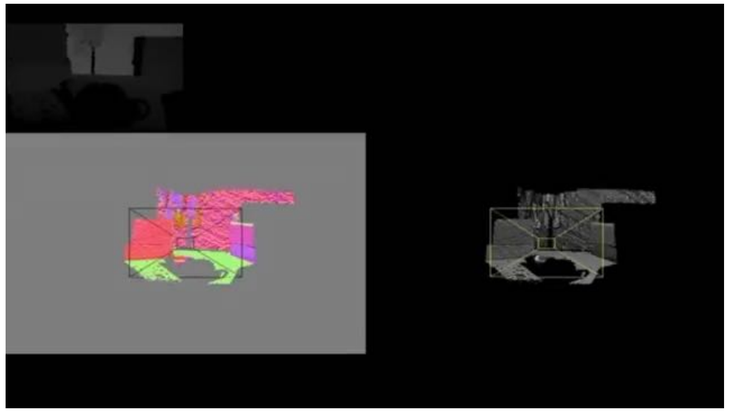
Credit: ARCore Depth API

- **Real-time dense surface scene reconstruction is the key to**
  - A full 3D perception of our environment
  - An important prerequisite for more realistic AR effects
    - Occlusion
    - Collision
    - Shadow mapping

Background

Current real-time dense reconstruction systems

With a depth sensor



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

With a monocular camera

Background

Current real-time dense reconstruction systems

With a depth sensor

Expensive!

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

With a monocular camera

Background

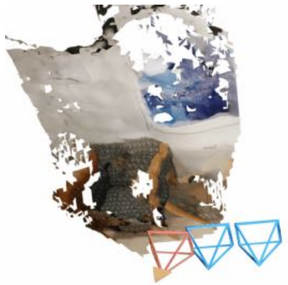
Current real-time dense reconstruction systems

With a depth sensor

Expensive!

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

With a monocular camera



Liu et al., "**Neural RGB→D Sensing: Depth and Uncertainty from a Video Camera**", *CVPR 2019*
Long et al., "**Occlusion-Aware Depth Estimation with Adaptive Normal Constraints**", *ECCV 2020*
Hou et al., "**Multi-View Stereo by Temporal Nonparametric Fusion**", *ICCV 2019*
Wang et al., "**MVDepthNet: Real-time Multiview Depth Estimation Neural Network**", *3DV 2018*
Yao et al., "**MVSNet: Depth Inference for Unstructured Multi-view Stereo**", *ECCV 2018*

Background

Current real-time dense reconstruction systems

With a depth sensor

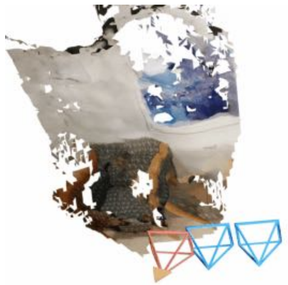
Expensive!

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

With a monocular camera

Redundant computation

Either layered or scattered results



Liu et al., "**Neural RGB→D Sensing: Depth and Uncertainty from a Video Camera**", *CVPR 2019*
Long et al., "**Occlusion-Aware Depth Estimation with Adaptive Normal Constraints**", *ECCV 2020*
Hou et al., "**Multi-View Stereo by Temporal Nonparametric Fusion**", *ICCV 2019*
Wang et al., "**MVDepthNet: Real-time Multiview Depth Estimation Neural Network**", *3DV 2018*
Yao et al., "**MVSNet: Depth Inference for Unstructured Multi-view Stereo**", *ECCV 2018*

Background

Current real-time dense reconstruction systems

With a depth sensor

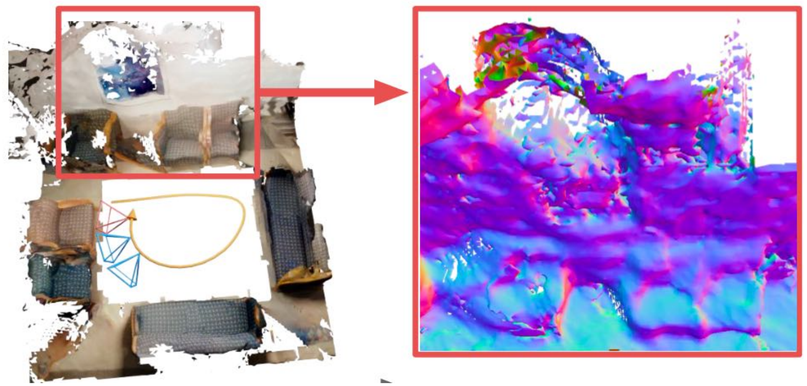
Expensive!

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

With a monocular camera

Redundant computation

Either layered or scattered results



Liu et al., "**Neural RGB→D Sensing: Depth and Uncertainty from a Video Camera**", *CVPR 2019*
Long et al., "**Occlusion-Aware Depth Estimation with Adaptive Normal Constraints**", *ECCV 2020*
Hou et al., "**Multi-View Stereo by Temporal Nonparametric Fusion**", *ICCV 2019*
Wang et al., "**MVDepthNet: Real-time Multiview Depth Estimation Neural Network**", *3DV 2018*
Yao et al., "**MVSNet: Depth Inference for Unstructured Multi-view Stereo**", *ECCV 2018*

Background

Current real-time dense reconstruction systems

With a depth sensor

With a monocular camera

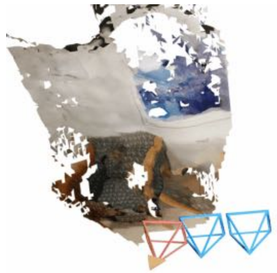
How to design a real-time and high-accuracy method for RGB scene reconstruction...

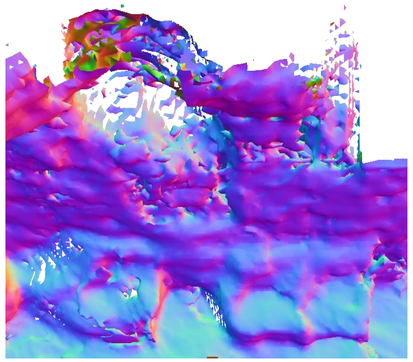
Motivation

Current real-time dense reconstruction systems

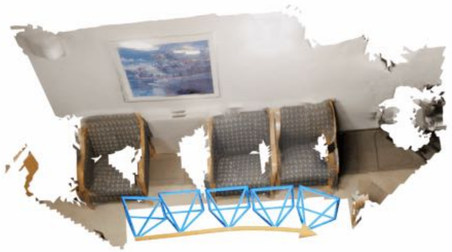
With a monocular camera

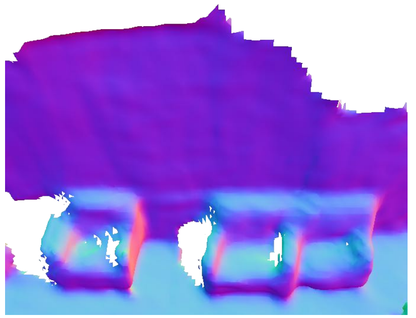
- Redundant computation
- Either layered or scattered results





Our solutions



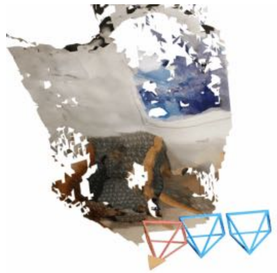


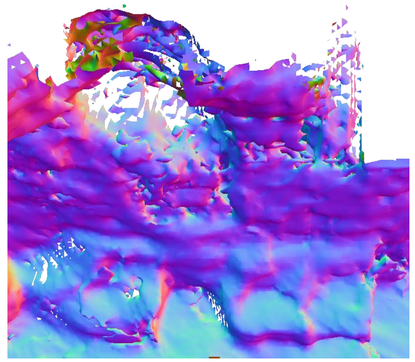
Motivation

Current real-time dense reconstruction systems

With a monocular camera

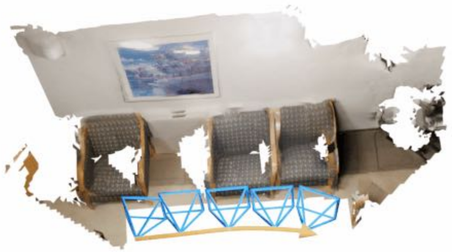
- Redundant computation
- Either layered or scattered results

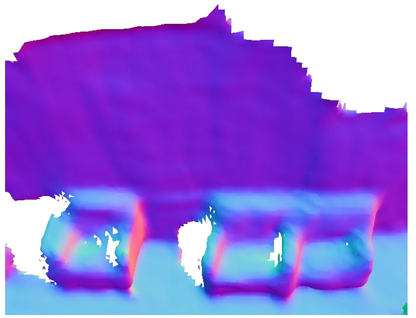




Our solutions

- Directly reconstruct local surfaces for each video fragment sequentially



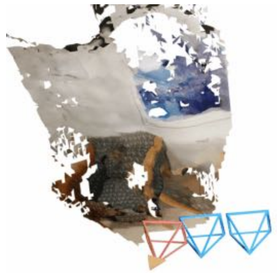


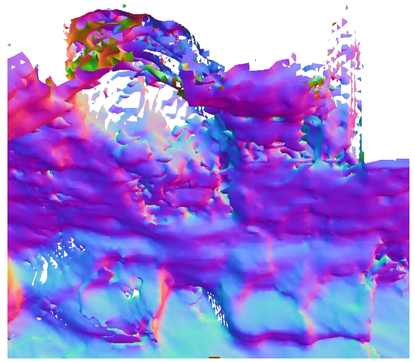
Motivation

Current real-time dense reconstruction systems

With a monocular camera

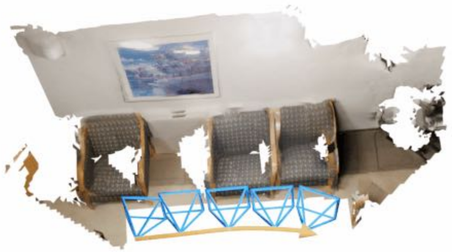
- Redundant computation
- Either layered or scattered results

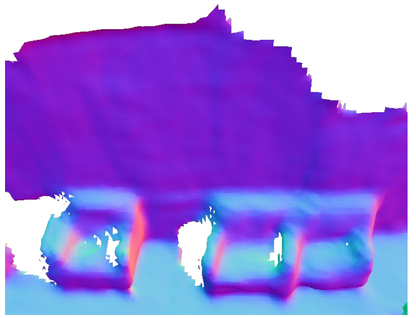




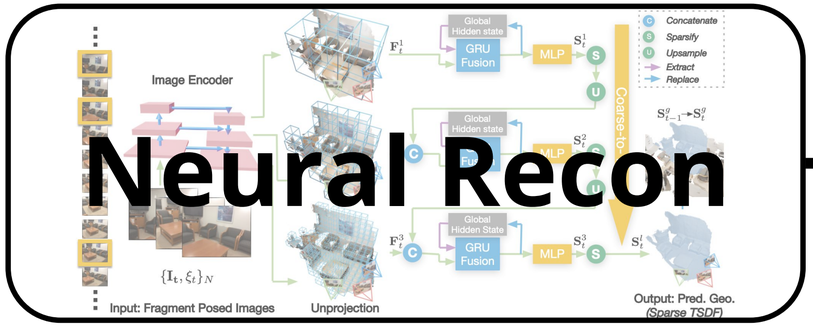
Our solutions

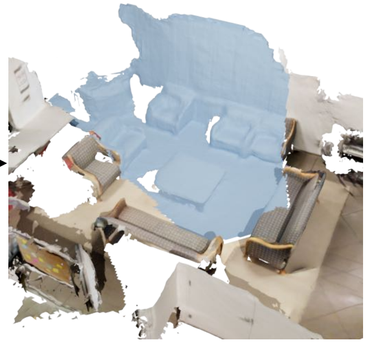
- Directly reconstruct local surfaces for each video fragment sequentially
- Learning-based TSDF fusion module





NeuralRecon



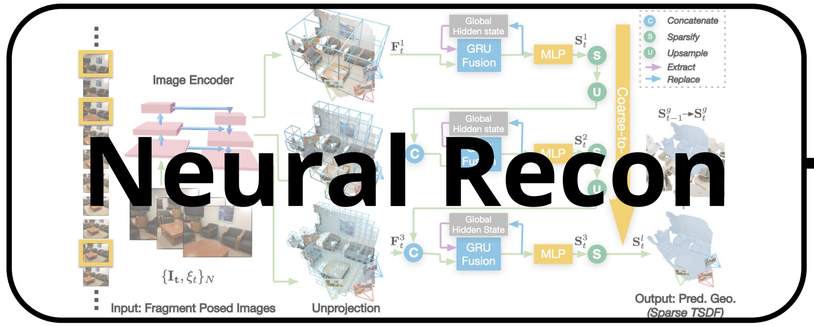


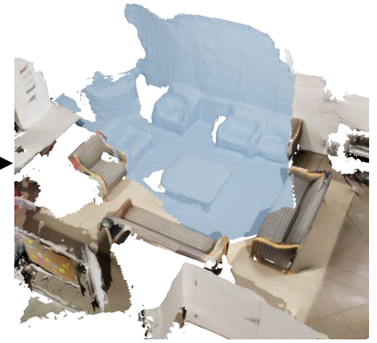
Input: Posed Images

End-to-End System

Output: Geometry
(Sparse TSDF)

NeuralRecon

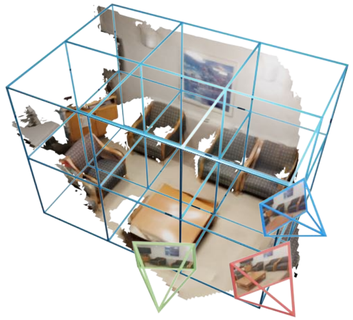




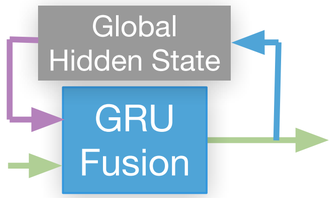
Input: Posed Images

End-to-End System

Output: Geometry
(Sparse TSDF)



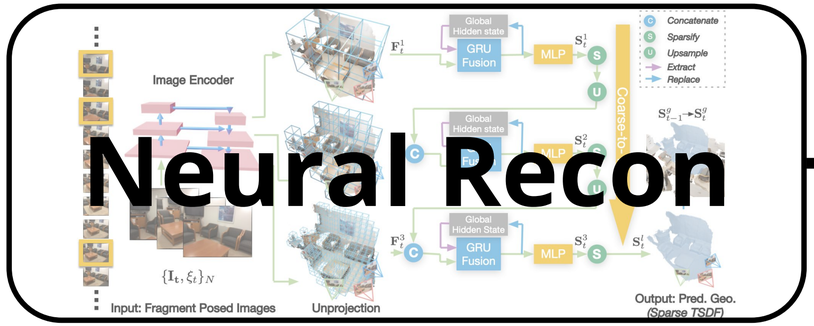
View-Independent
Volume



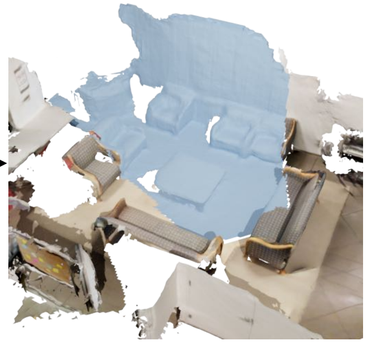
Learning-based
TSDF Fusion

NeuralRecon

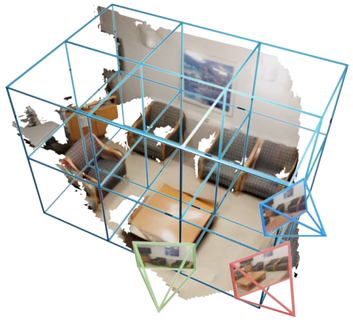
Input: Posed Images



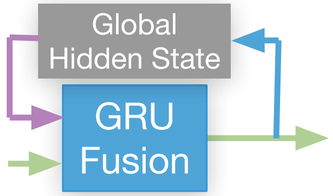
End-to-End System



Output: Geometry
(Sparse TSDF)

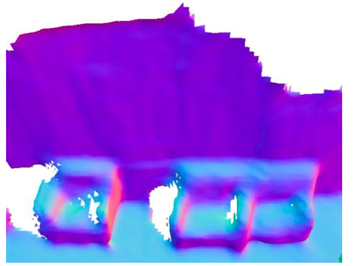


View-Independent Volume



Learning-based TSDF Fusion

Real Time



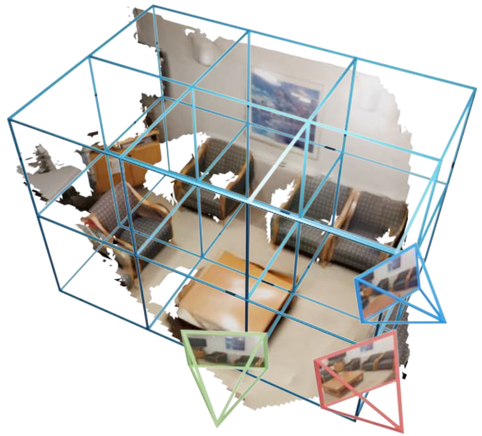
Coherent

Fragment Reconstruction

Input: Posed Images

Fragment Reconstruction

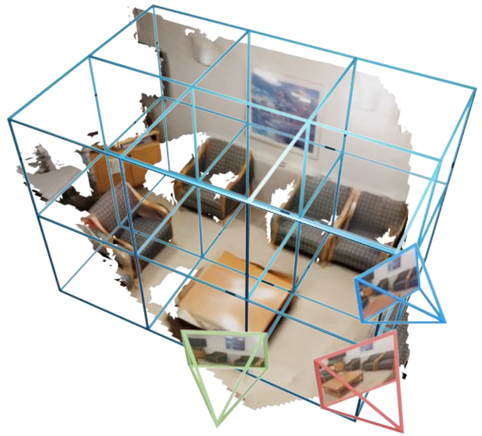
Input: Posed Images



View-Independent Volume

Fragment Reconstruction

Input: Posed Images

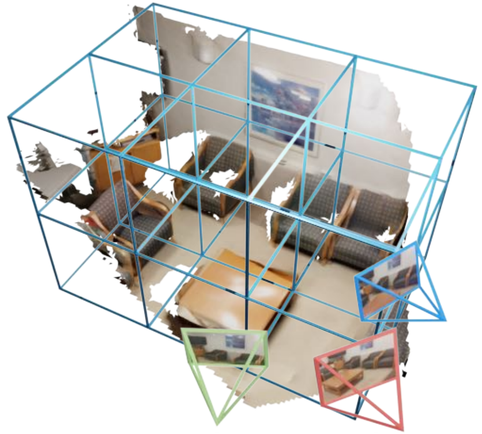


View-Independent Volume

Sparse Conv

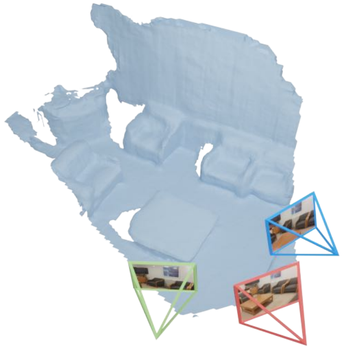
Fragment Reconstruction

Input: Posed Images



View-Independent Volume

Sparse Conv

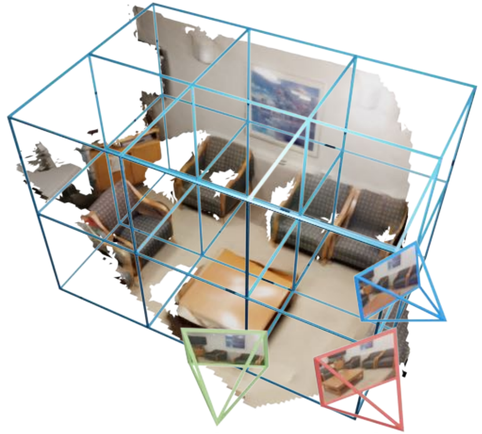


Fragment Geometry

(Sparse TSDF)

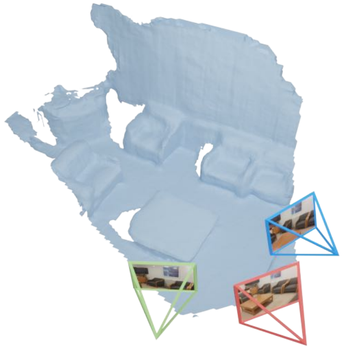
Fragment Reconstruction

Input: Posed Images



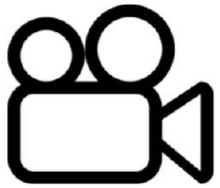
View-Independent Volume

Sparse Conv



Fragment Geometry

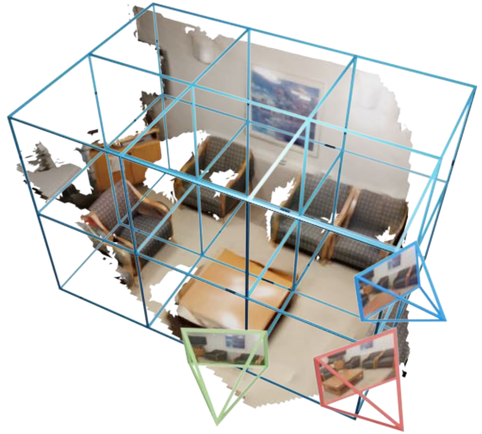
(Sparse TSDF)



Depth Sensor

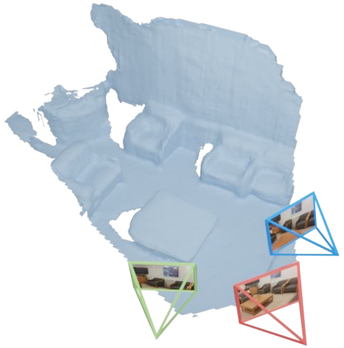
Fragment Reconstruction

Input: Posed Images



View-Independent Volume

Sparse Conv



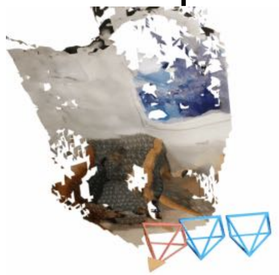
Fragment Geometry

(Sparse TSDF)

Why is it Better?

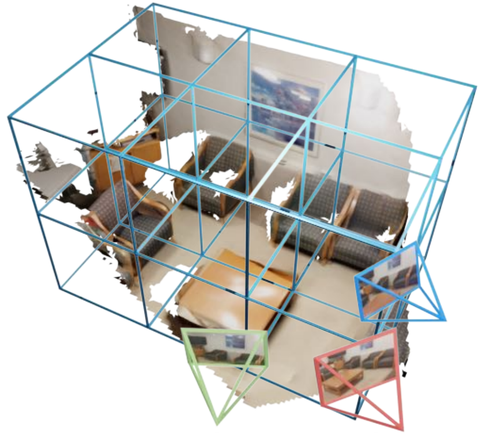
Volume-based

Depth-based



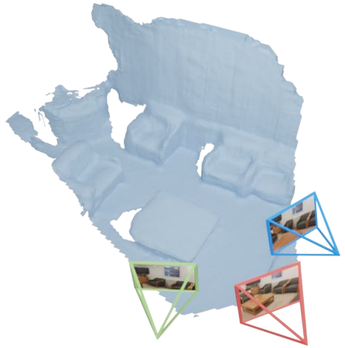
Fragment Reconstruction

Input: Posed Images



View-Independent Volume

Sparse Conv



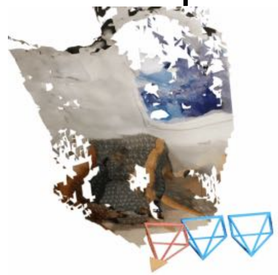
Fragment Geometry
(Sparse TSDF)

Why is it Better?

1. View-independent Volume
—> *locally coherent, faster*

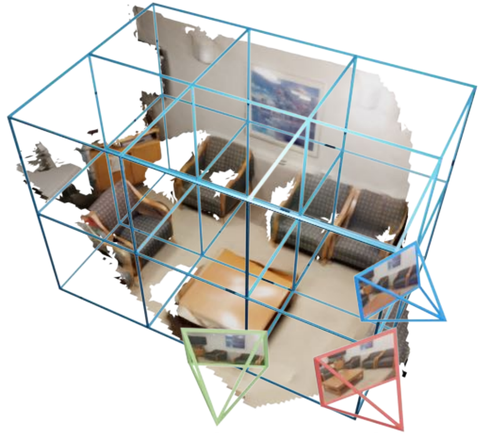
Volume-based

Depth-based



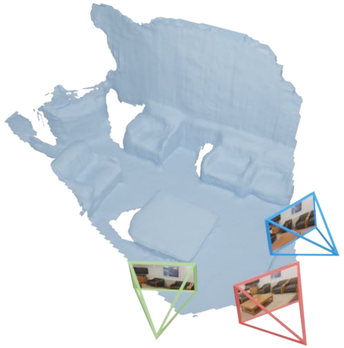
Fragment Reconstruction

Input: Posed Images



View-Independent Volume

Sparse Conv



Fragment Geometry
(Sparse TSDF)

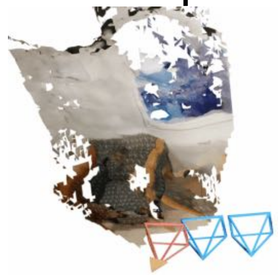
Why is it Better?

1. View-independent Volume
—> *locally coherent, faster*

2. Directly regress TSDF rather than an intermediate representation of depth maps
—> *faster*

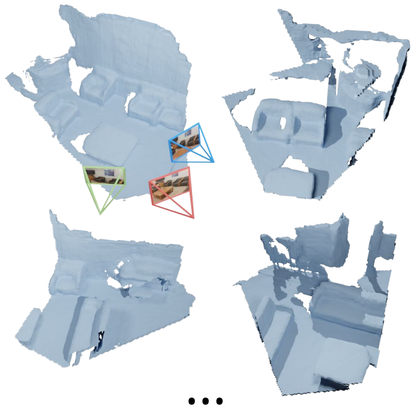
Volume-based

Depth-based



Fragment Fusion

Fragment Reconstruction



...

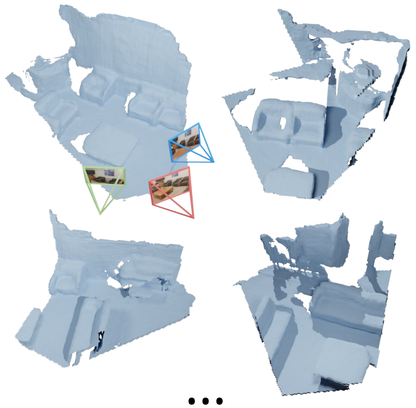
Fragment Geometry

(Sparse TSDF)

Depth Sensor

Fragment Fusion

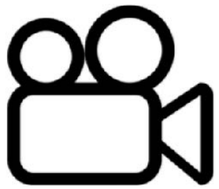
Fragment Reconstruction



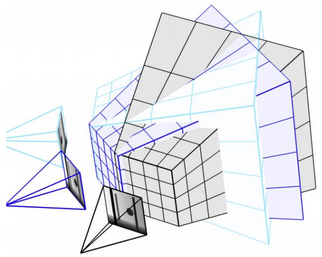
...

Fragment Geometry

(Sparse TSDF)



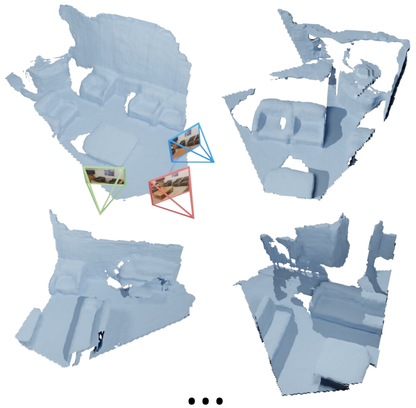
Depth Sensor



Fusion

Fragment Fusion

Fragment Reconstruction



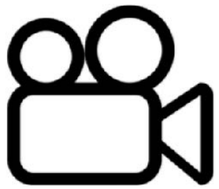
...

Fragment Geometry

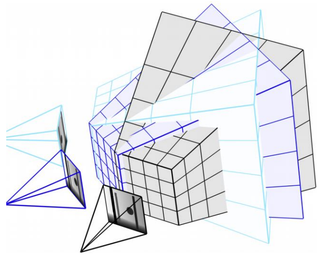
(Sparse TSDF)

?

TSDF Fusion



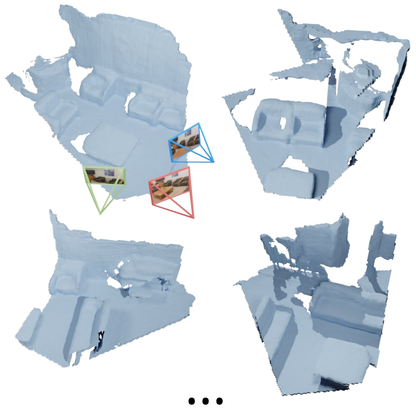
Depth Sensor



Fusion

However...

Fragment Reconstruction



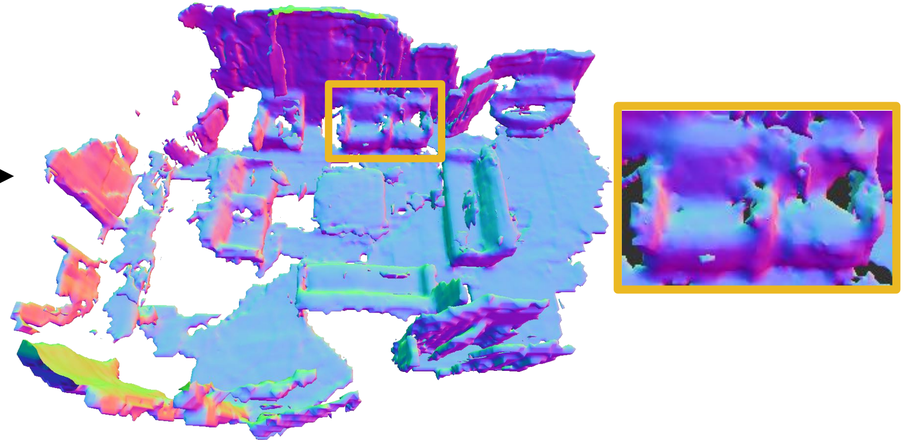
...

Fragment Geometry

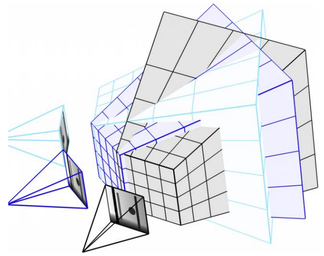
(Sparse TSDF)

?

TSDF Fusion



Depth Sensor

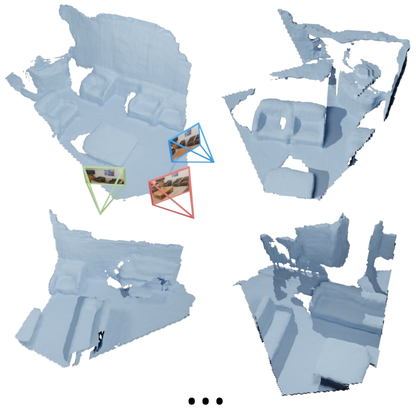


Fusion

Too many artifacts!

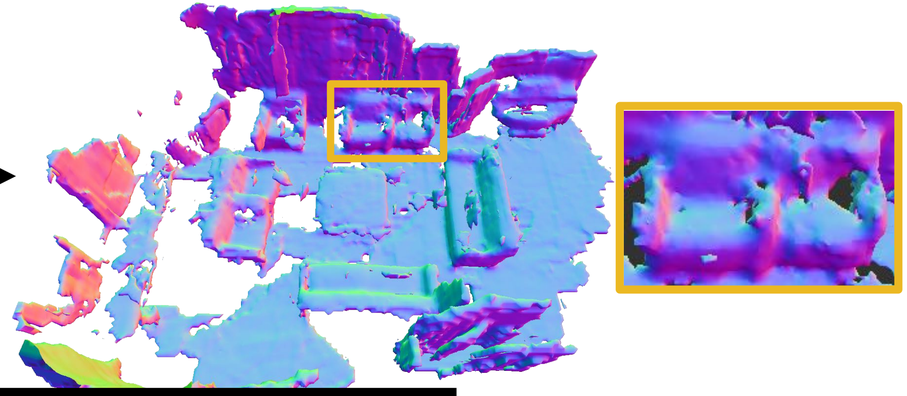
However…

Fragment Reconstruction



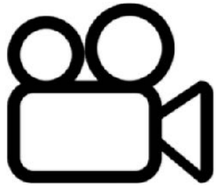
?

TSDF Fusion

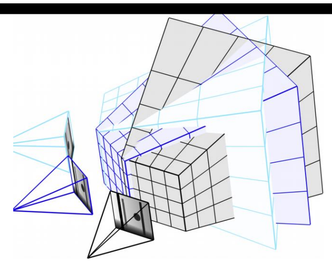


Model what we can and learn what we can't!

…ny artifacts!



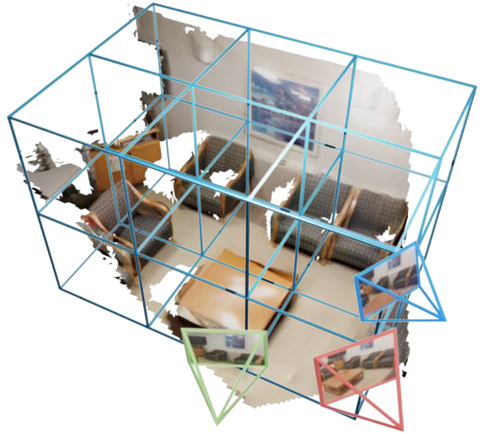
Depth Sensor



Fusion

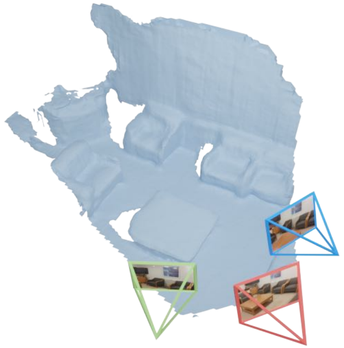
Improving Fusion

Input: Posed Images



View-Independent Volume

Sparse Conv

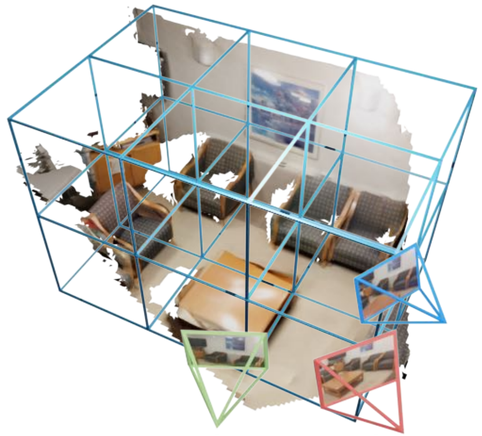


Fragment Geometry

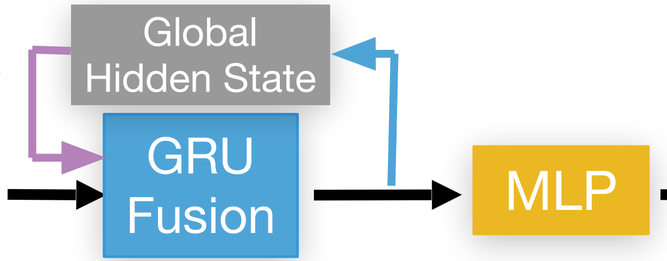
(Sparse TSDF)

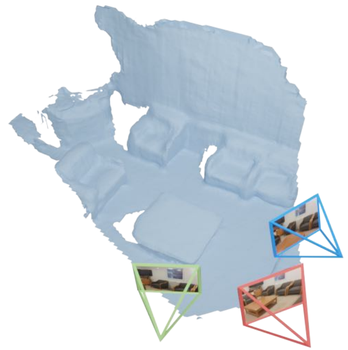
Improving Fusion

Input: Posed Images



View-Independent Volume



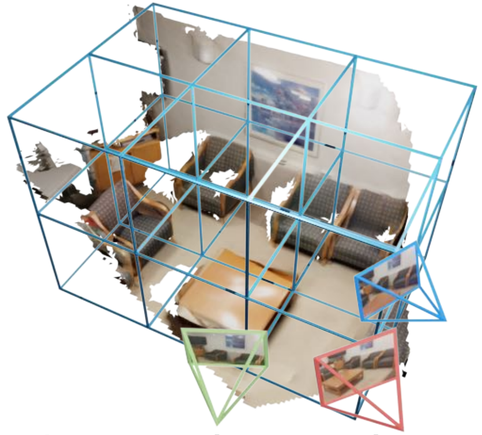


Fragment Geometry

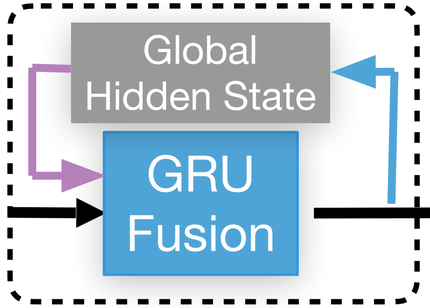
(Sparse TSDF)

Joint Estimation and Fusion

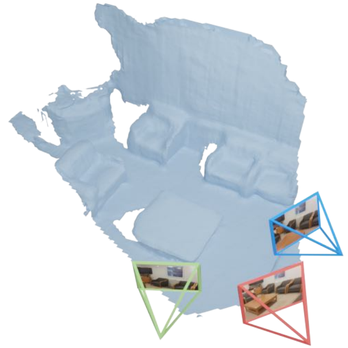
Input: Posed Images



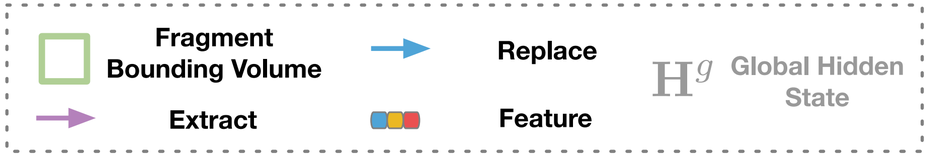
View-Independent Volume

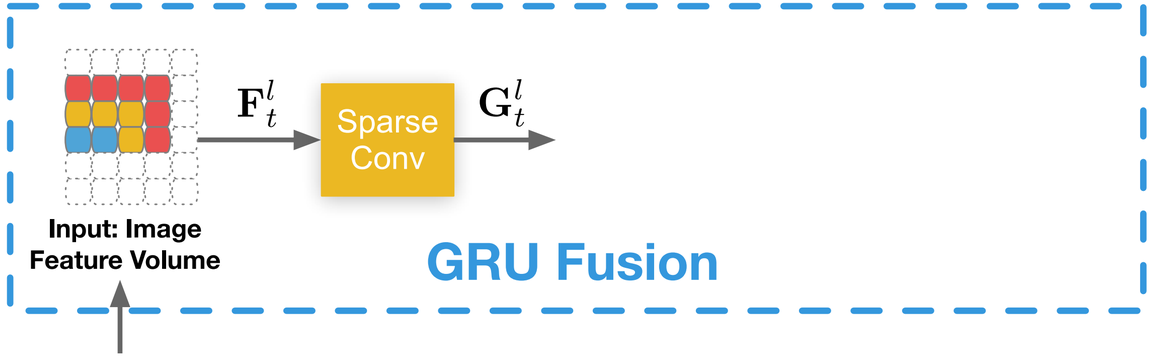


MLP



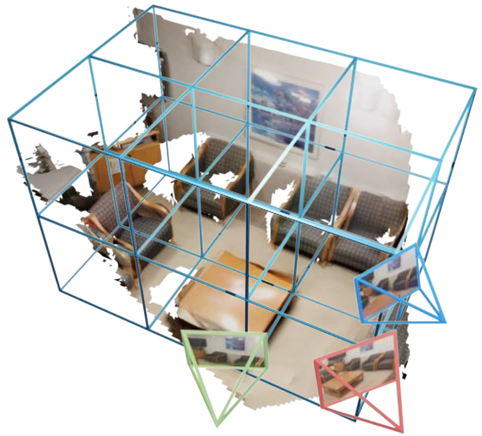
Fragment Geometry
(Sparse TSDF)



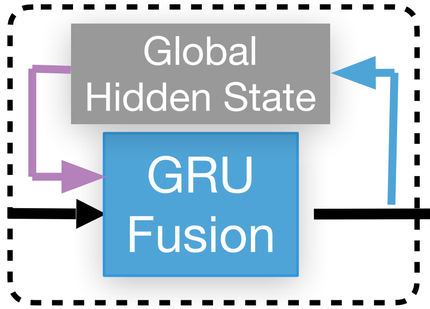


Joint Estimation and Fusion

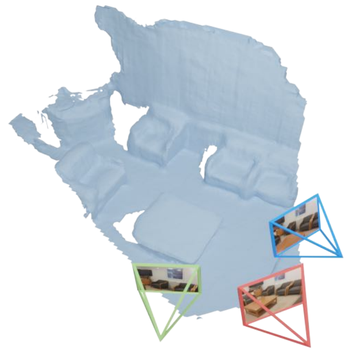
Input: Posed Images



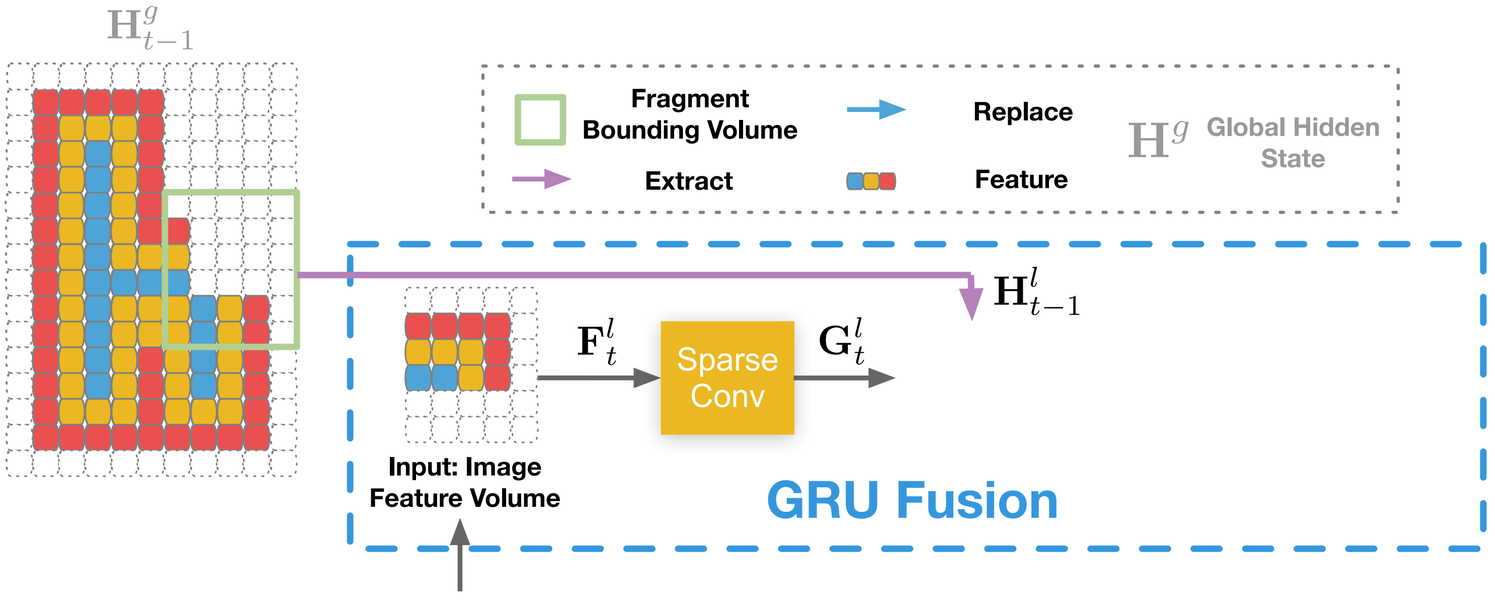
View-Independent Volume



MLP

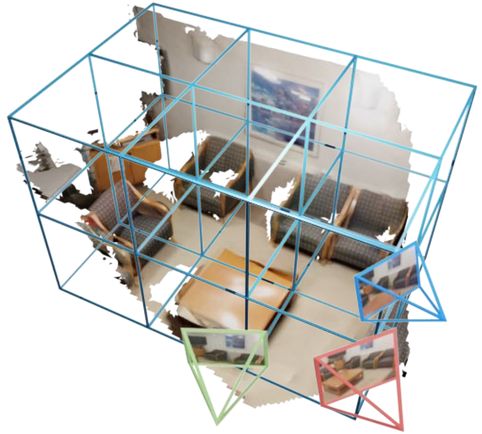


Fragment Geometry
(Sparse TSDF)

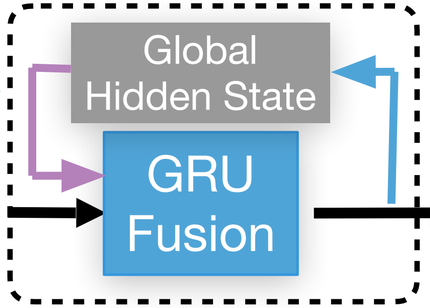


Joint Estimation and Fusion

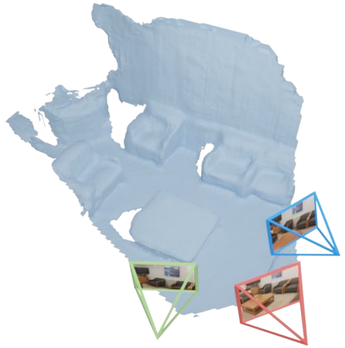
Input: Posed Images



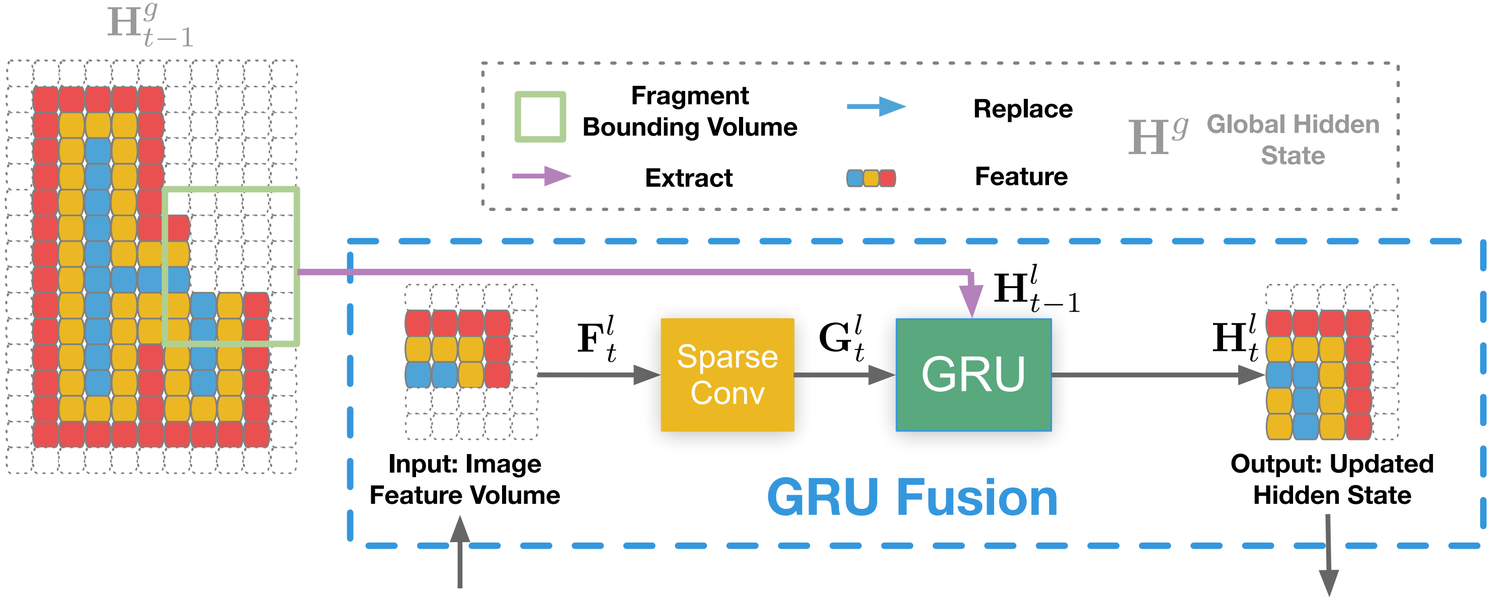
View-Independent Volume



MLP

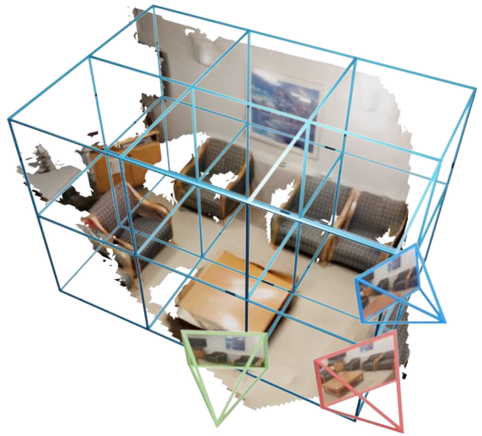


Fragment Geometry
(Sparse TSDF)

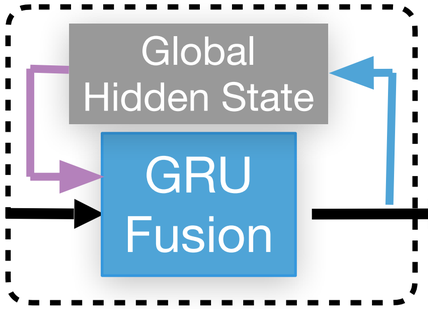


Joint Estimation and Fusion

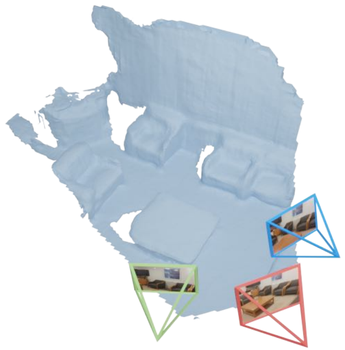
Input: Posed Images



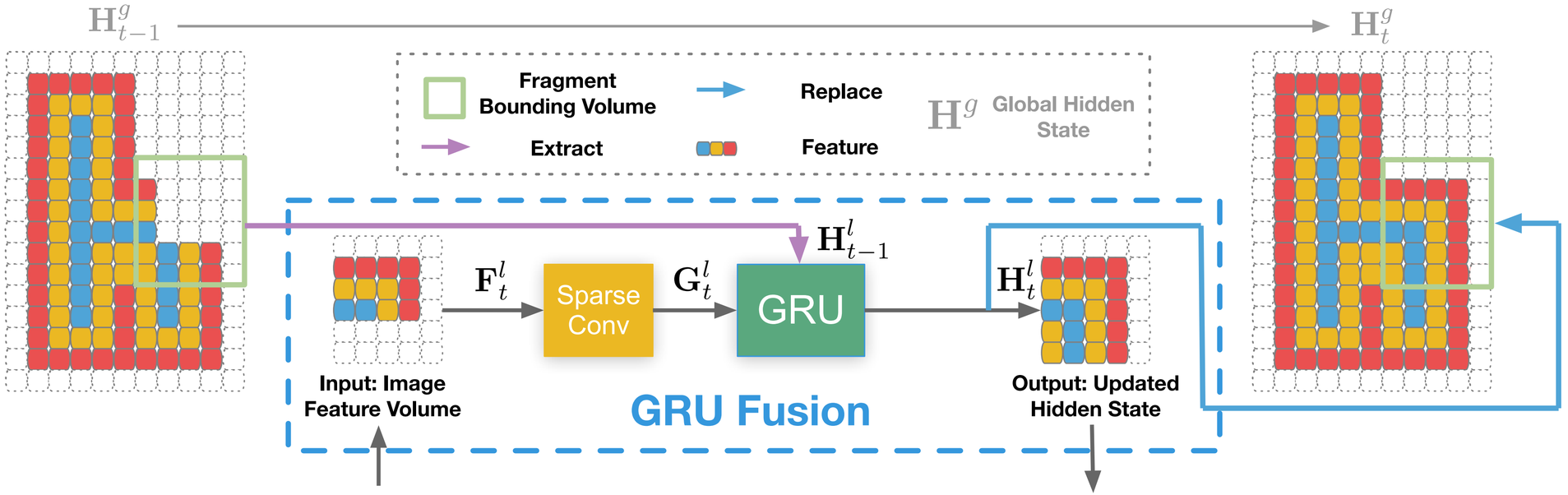
View-Independent Volume



MLP

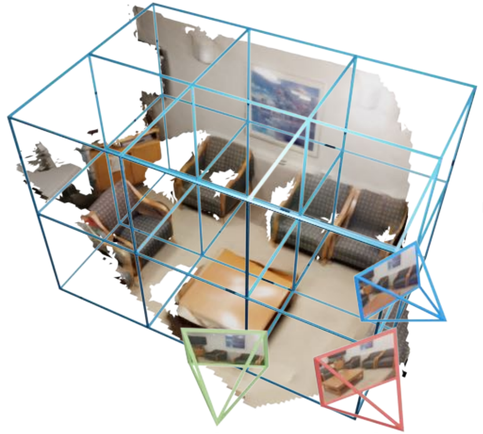


Fragment Geometry
(Sparse TSDF)

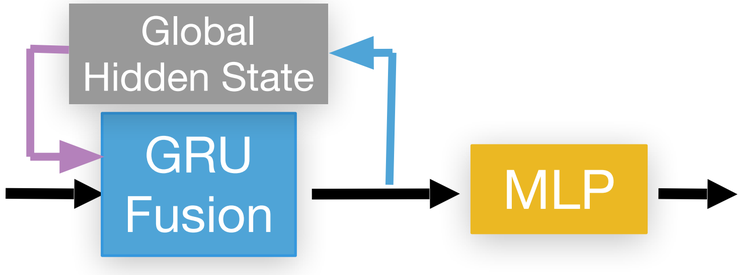


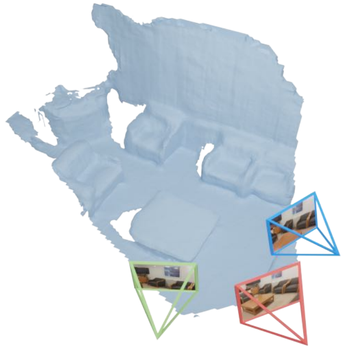
Joint Estimation and Fusion

Input: Posed Images



View-Independent Volume

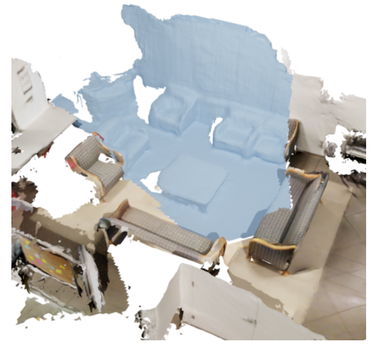




Fragment Geometry

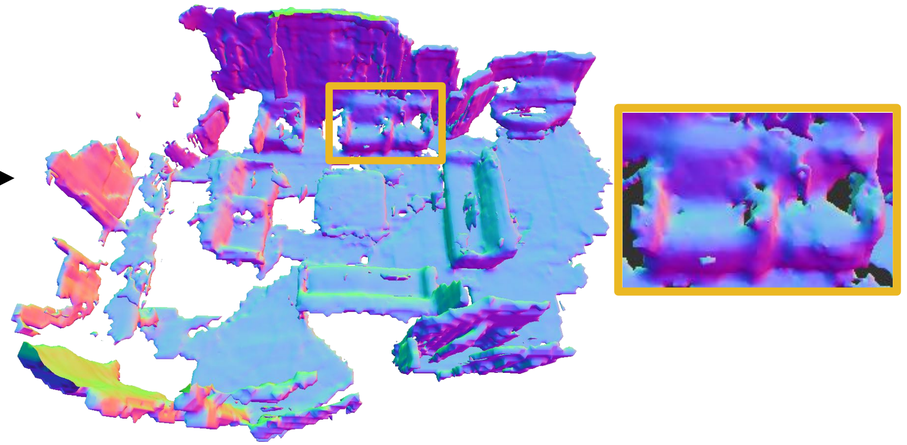
(Sparse TSDF)

Replace



Improving Fusion

Fragment Reconstruction → TSDF Fusion →

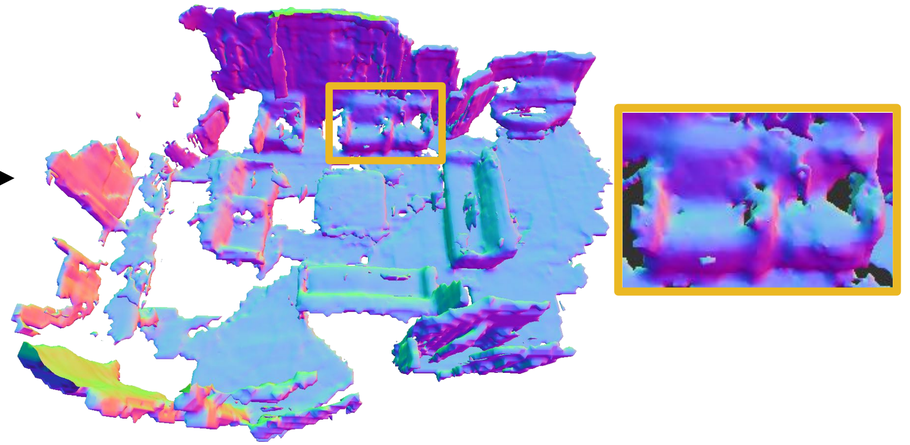


VS

Joint Estimation and Fusion →

Improving Fusion

Fragment Reconstruction → TSDF Fusion →



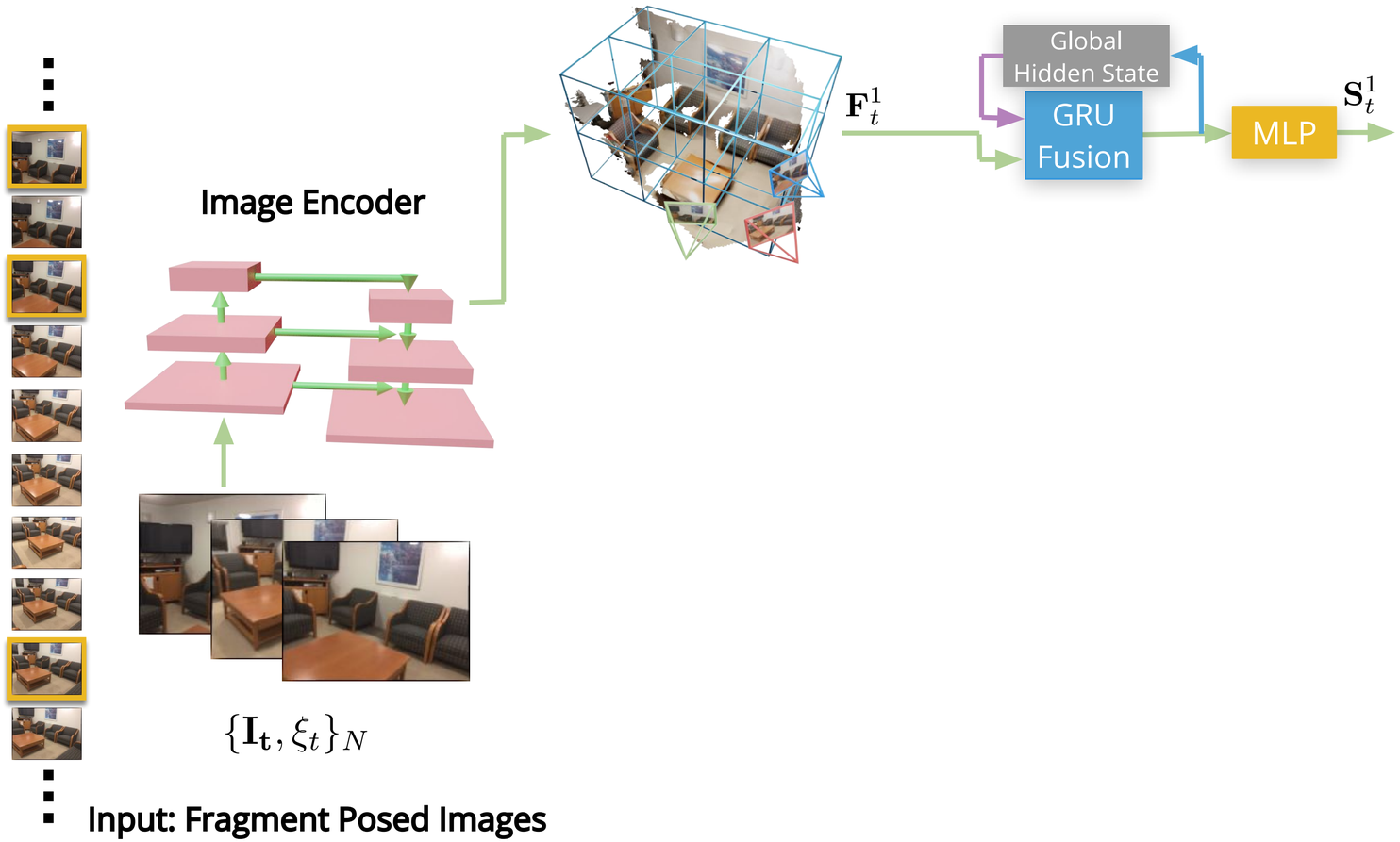
VS

Joint Estimation and Fusion →

More accurate!

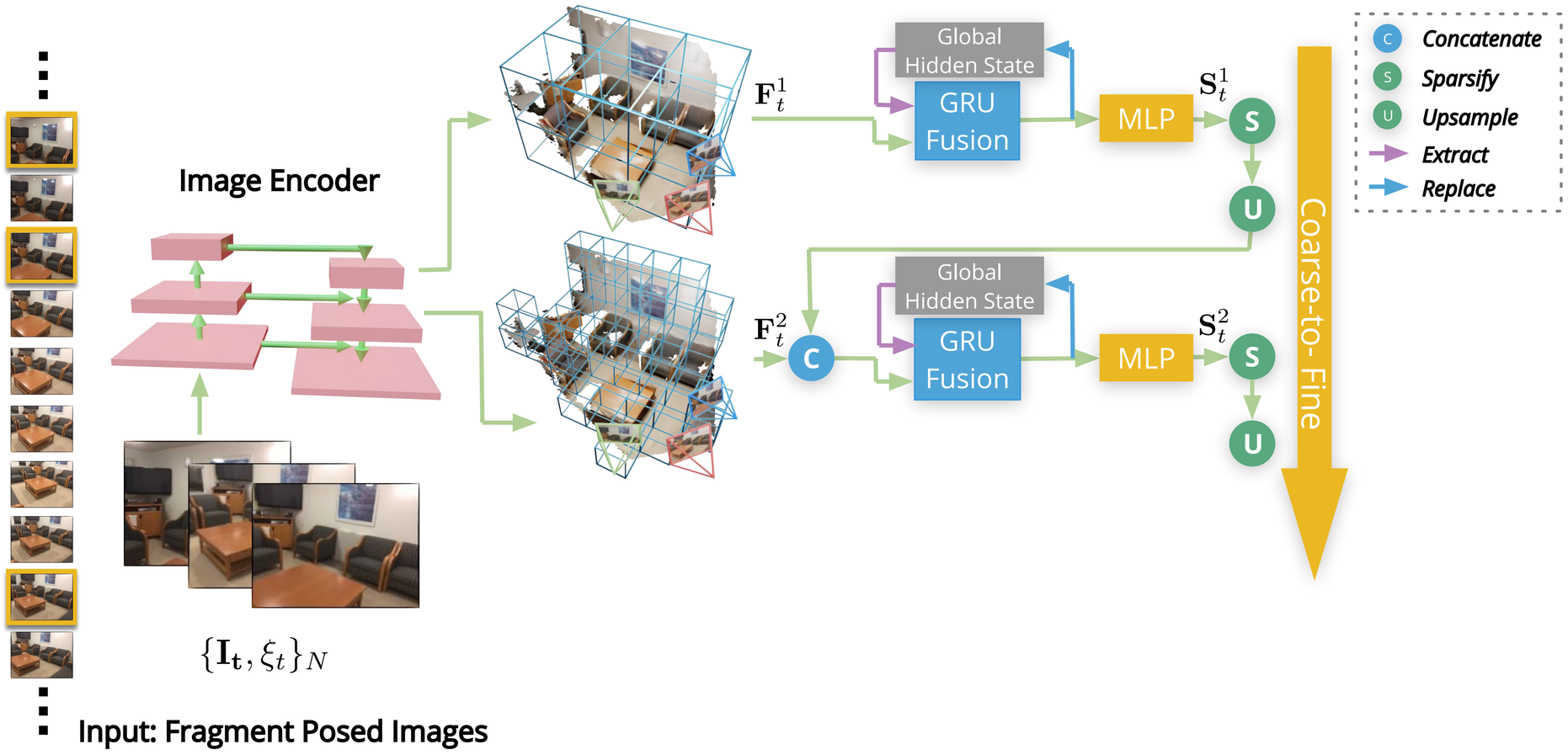
More coherent!

Network Architecture



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

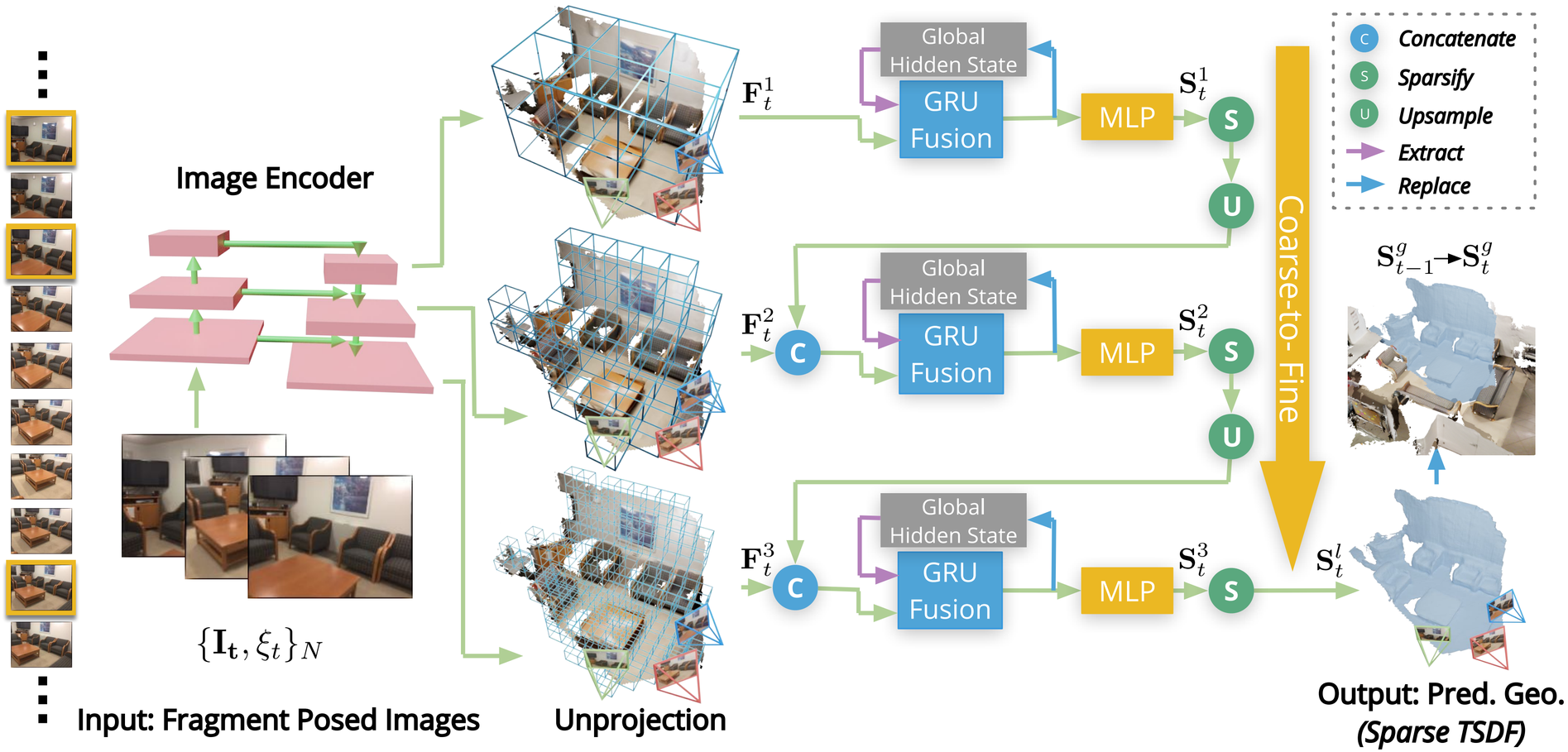
Network Architecture



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

S Filter by Occupancy Score > 0

Network Architecture

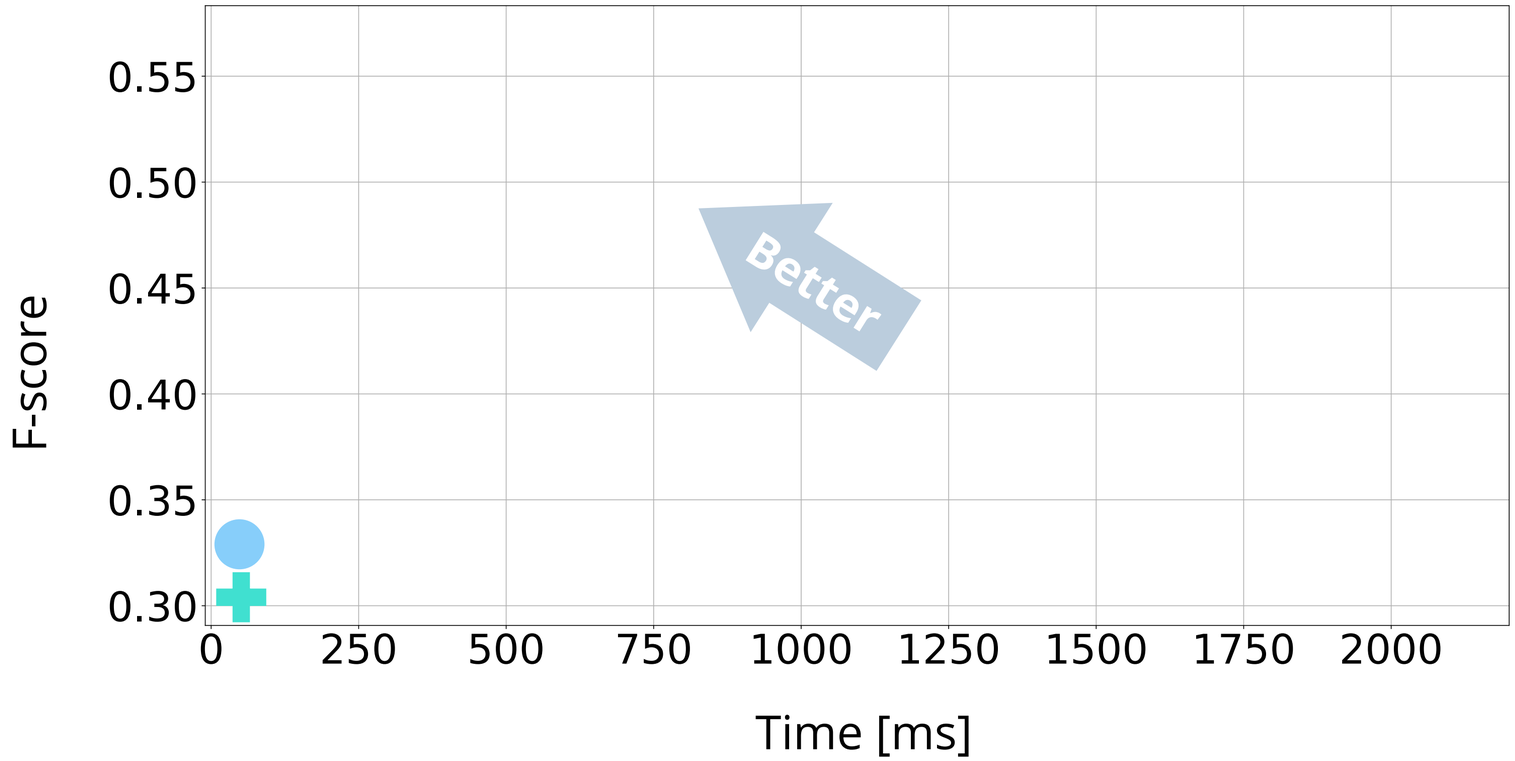


Output of *MLP* : **Occupancy Score** and **SDF**

S Filter by Occupancy Score > 0

Experiments

Quantitative Result

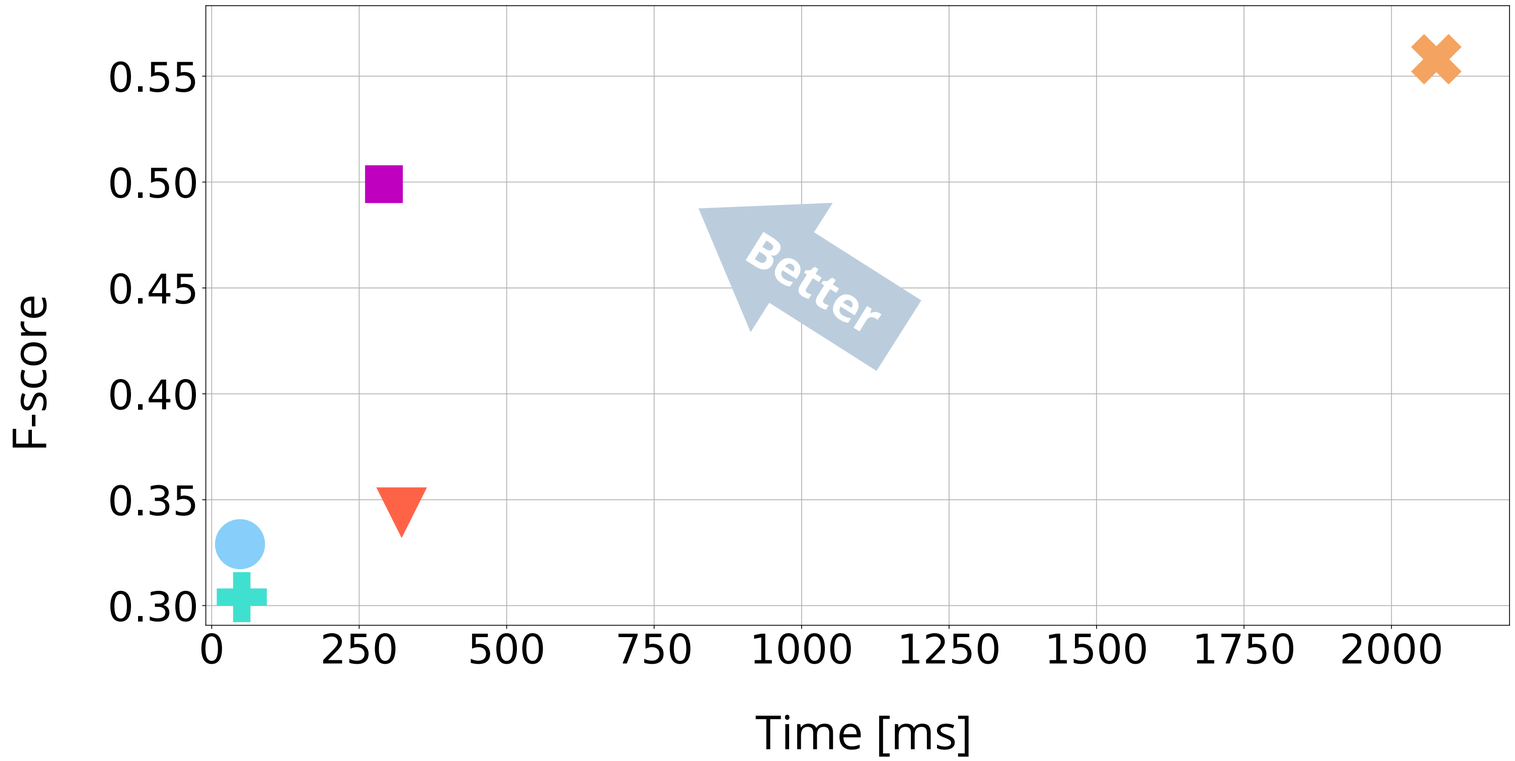


Real-time methods: MVDepthNet GPMVS

Fast **Accurate**

Experiments

Quantitative Result



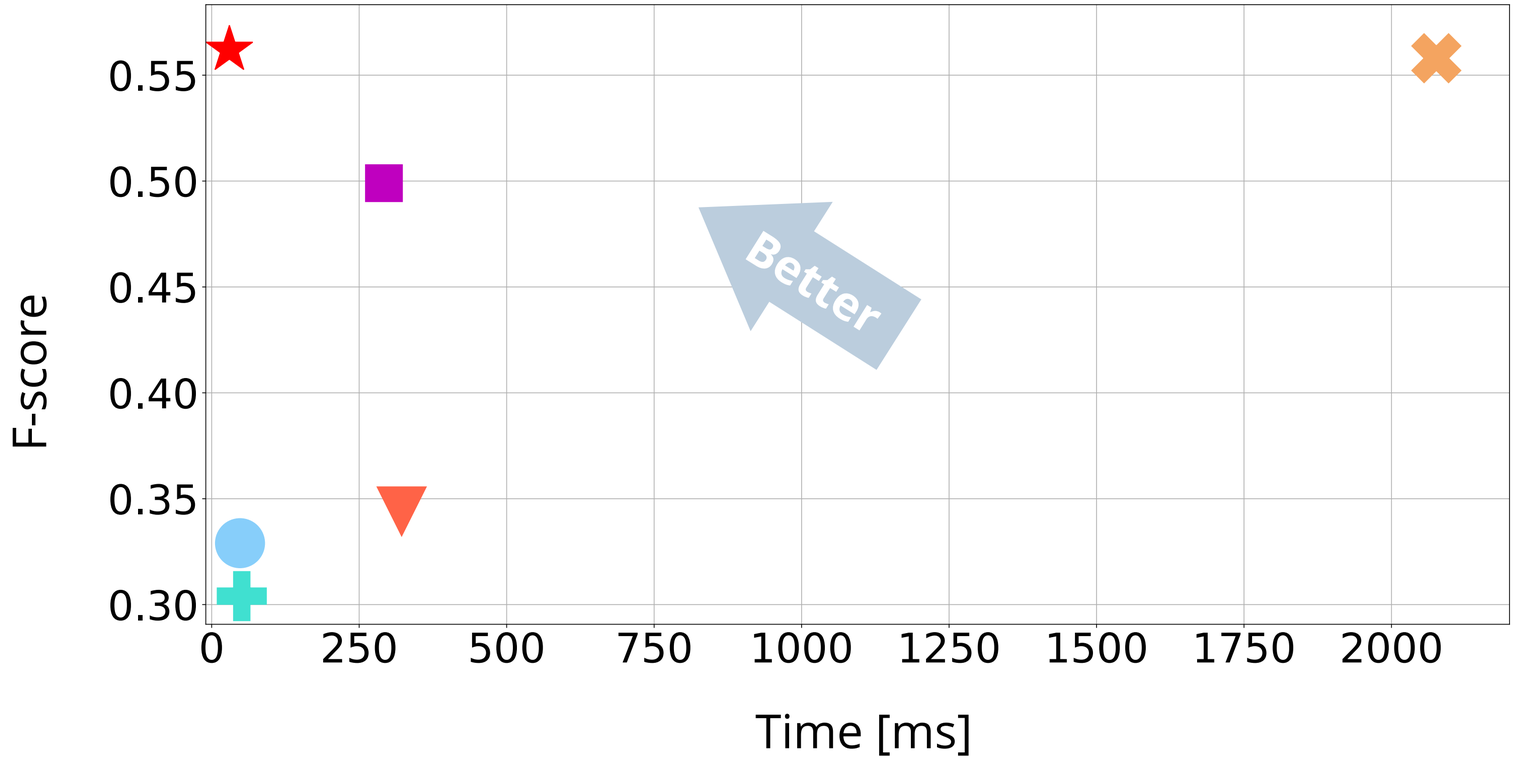
Real-time methods: MVDepthNet GPMVS

Multiple View Stereo methods: DPSNet COLMAP Atlas

Fast **Accurate**

Experiments

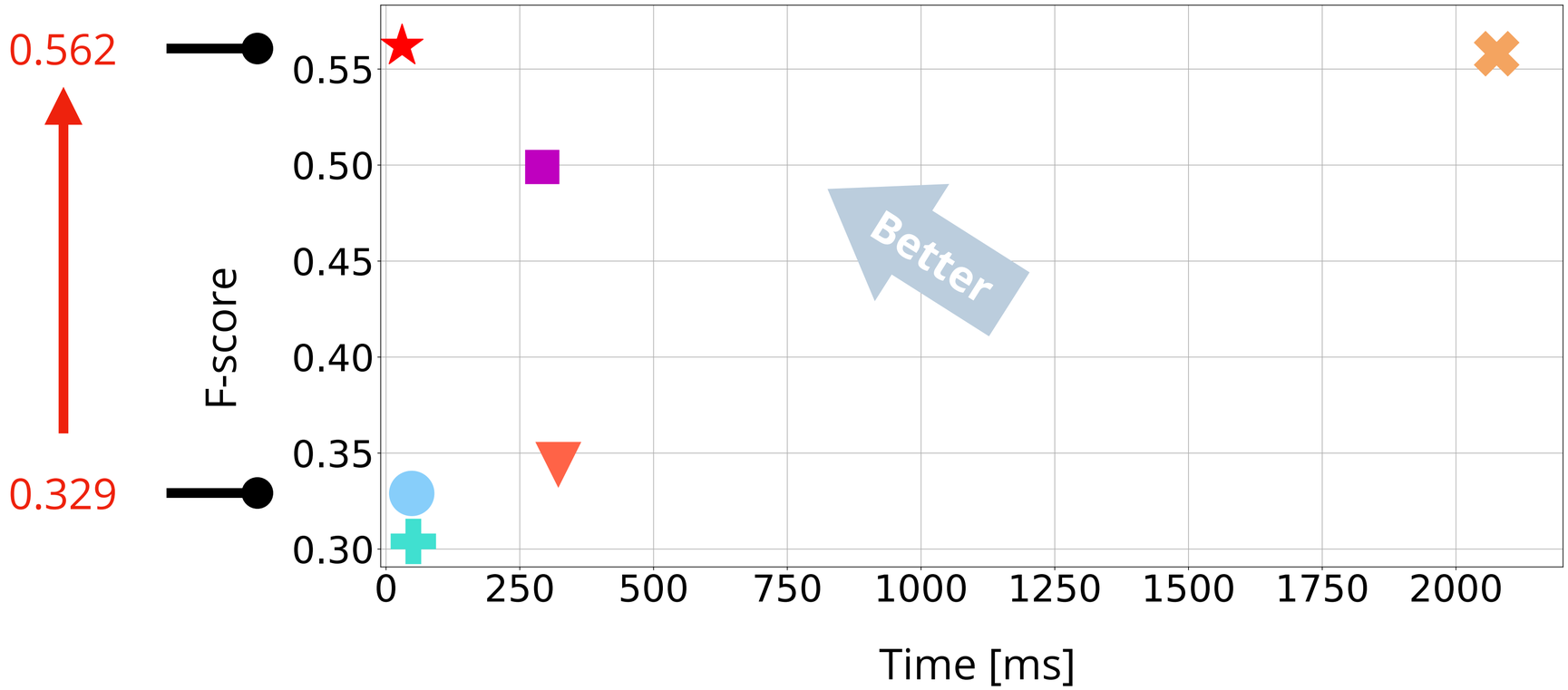
Quantitative Result



				Fast	Accurate
Real-time methods:	MVDepthNet	GPMVS		👍	👎
Multiple View Stereo methods:	DPSNet	COLMAP	Atlas	👎	👍
	Ours			👍	👍

Experiments

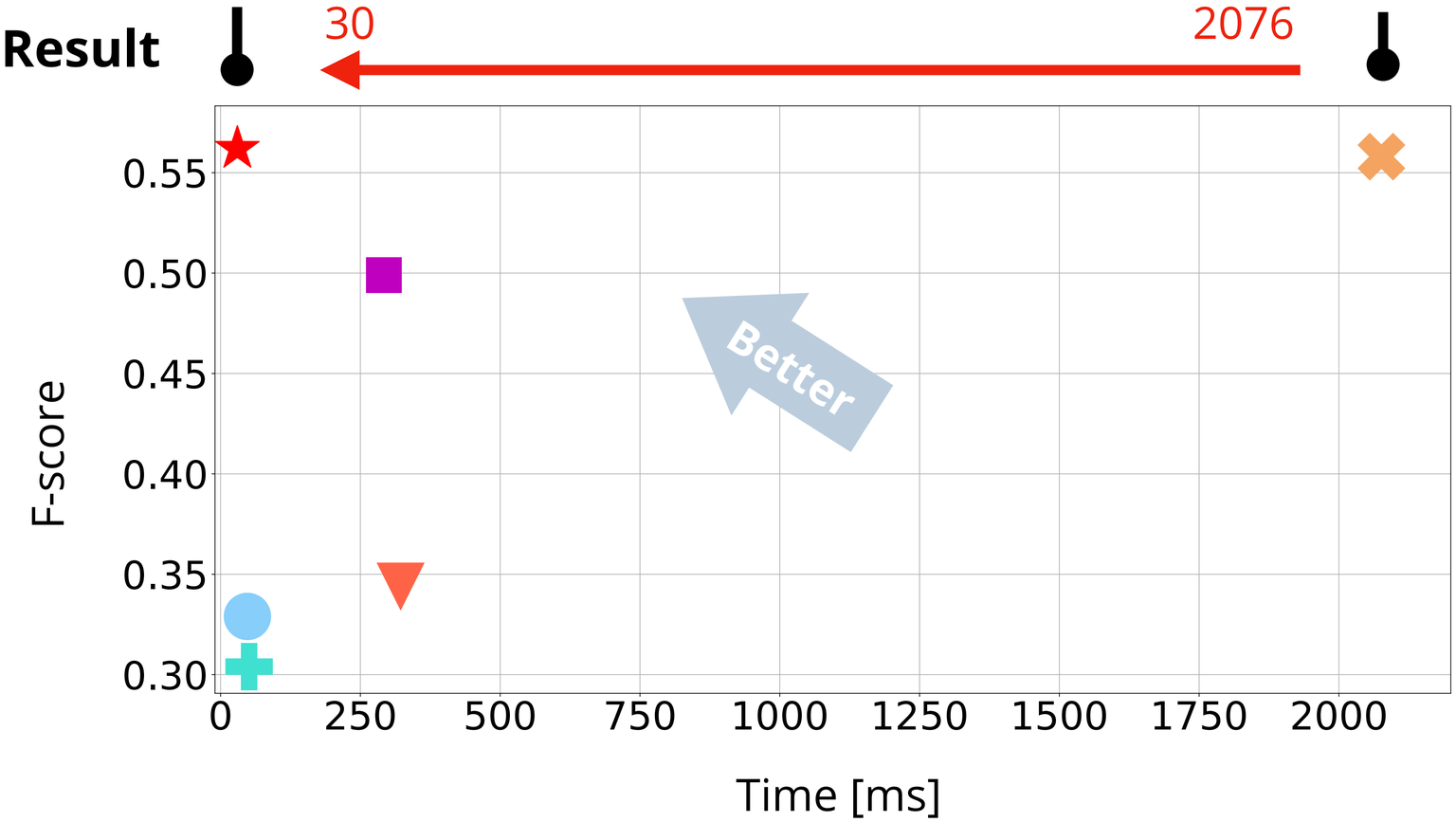
Quantitative Result



				Fast	Accurate
Real-time methods:	MVDepthNet	GPMVS		👍	👎
Multiple View Stereo methods:	DPSNet	COLMAP	Atlas	👎	👍
	Ours			👍	👍

Experiments

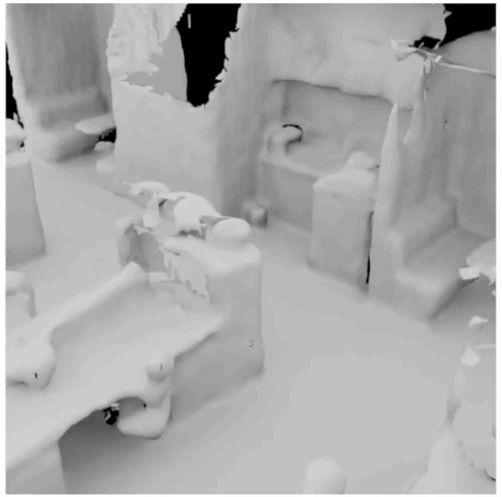
Quantitative Result



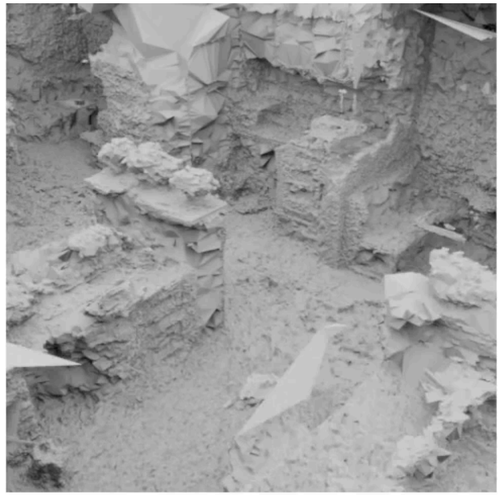
				Fast	Accurate
Real-time methods:	MVDepthNet	GPMVS		👍	👎
Multiple View Stereo methods:	DPSNet	COLMAP	Atlas	👎	👍
	Ours			👍	👍

Experiments

Qualitative Result — Scene 1

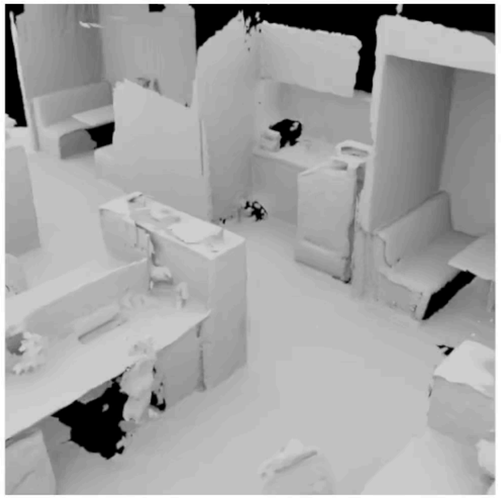


Ours (30ms)



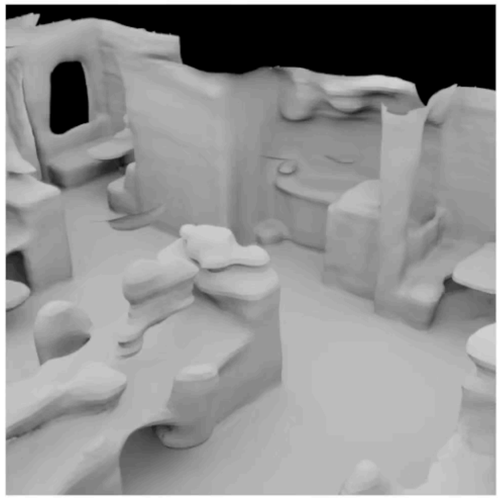
COLMAP (2076ms)

DeepV2D (347ms)



Ground Truth

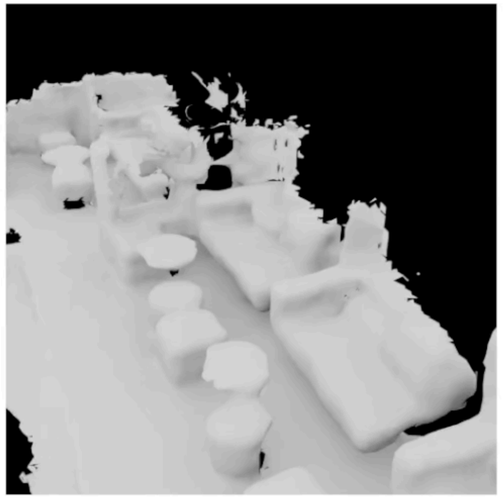
CNMNet (80ms)



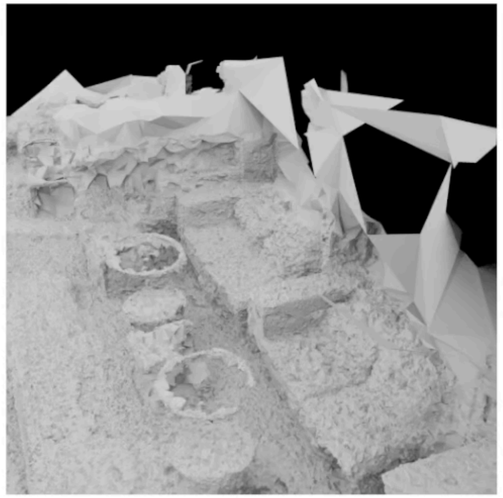
Atlas (292ms)

Experiments

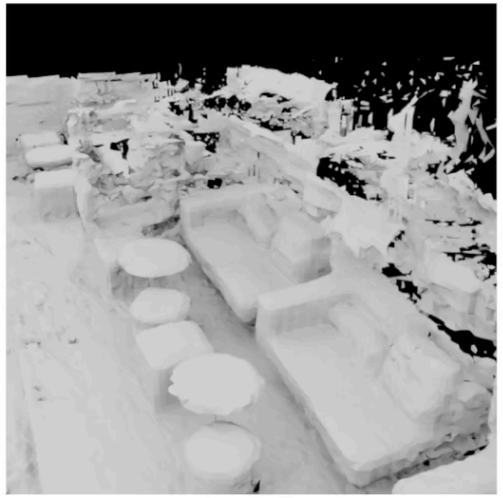
Qualitative Result — Scene 2



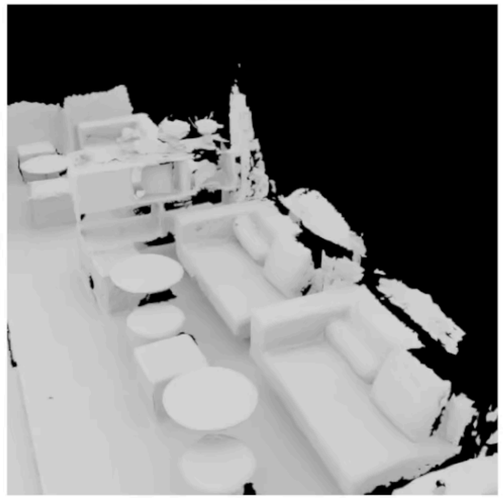
Ours (30ms)



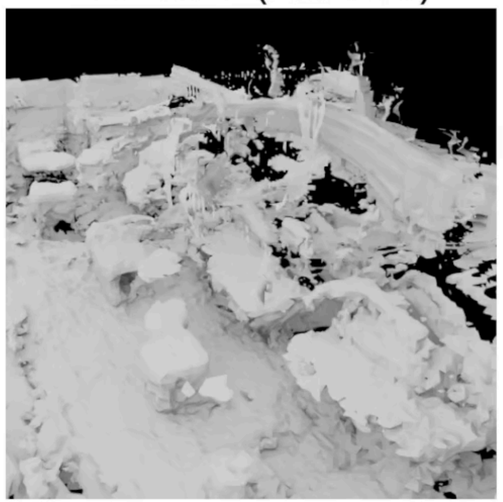
COLMAP (2076ms)



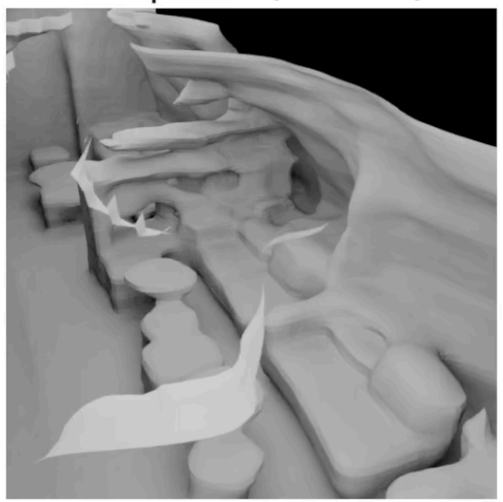
DeepV2D (347ms)



Ground Truth



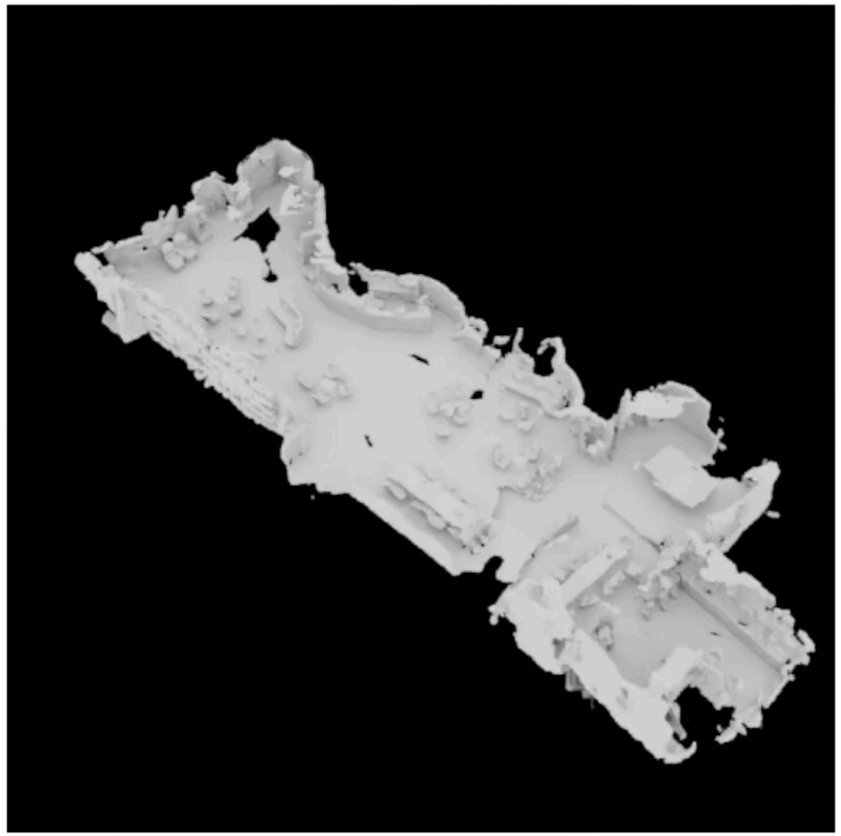
CNMNet (80ms)



Atlas (292ms)

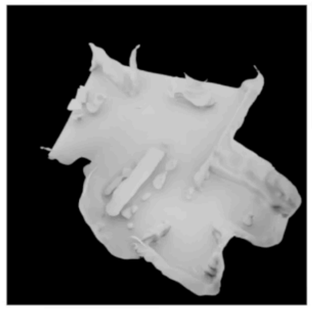
Experiments

Qualitative Result — Comparison with Atlas on a large scene (30m x 10m)



Ours

GPU Memory: 9GB



Atlas

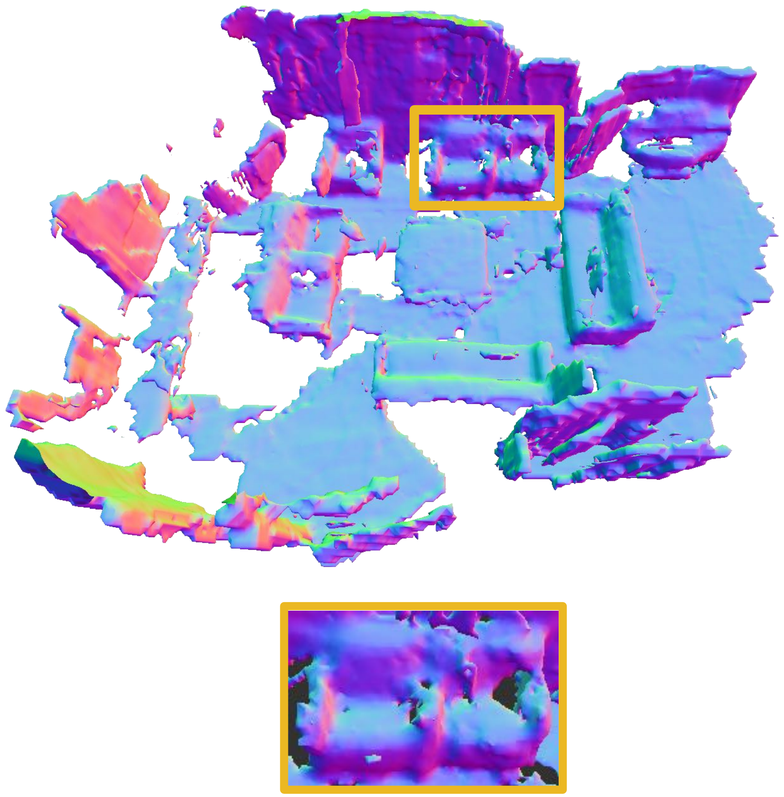
GPU Memory: OOM

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

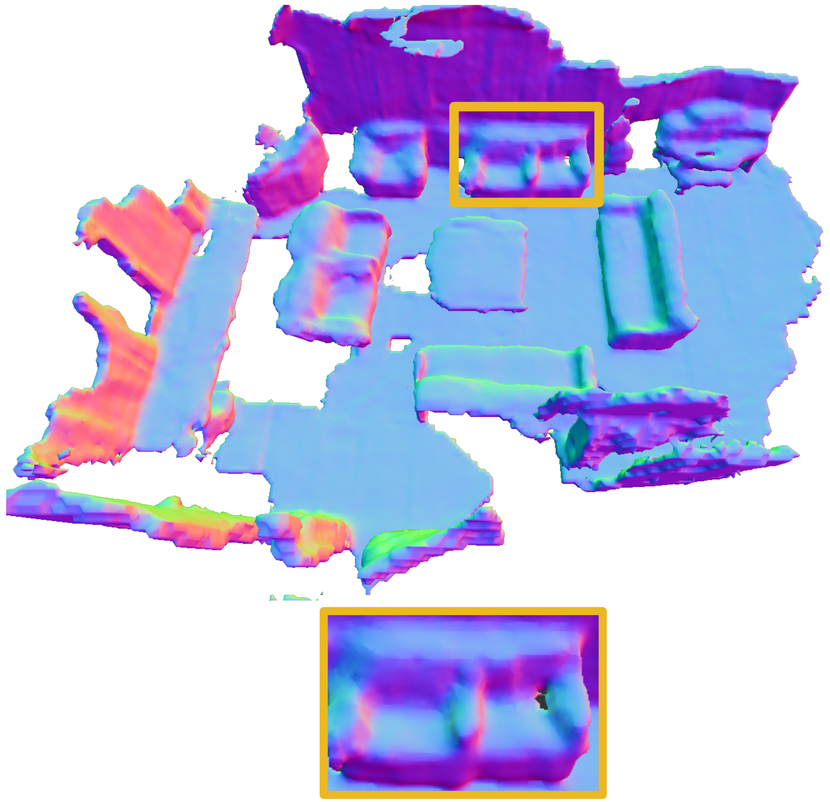
Ablation Study

Does Joint Estimation and Fusion Matter?

Fragment Reconstruction → TSDF Fusion

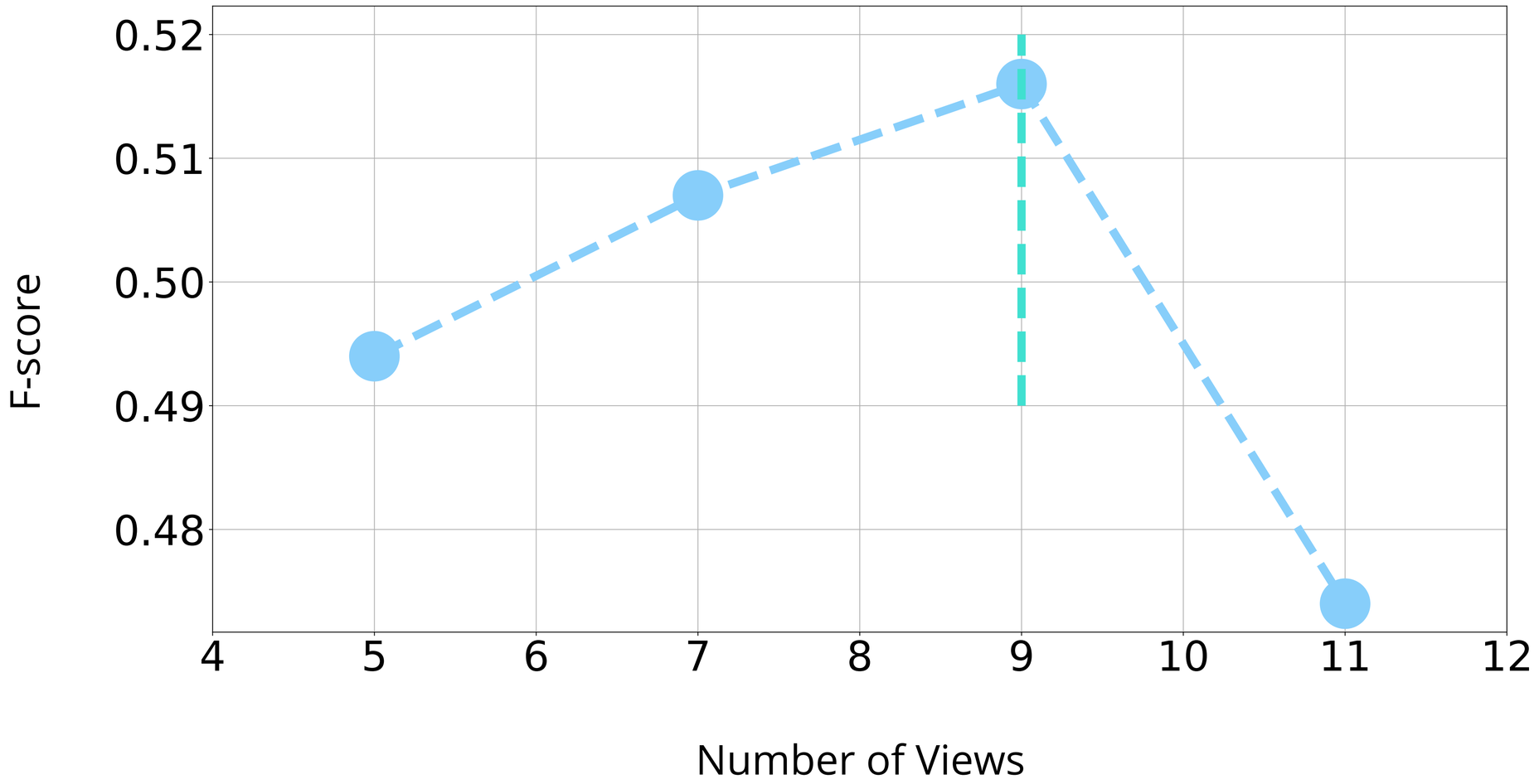


Joint Estimation and Fusion



Ablation Study

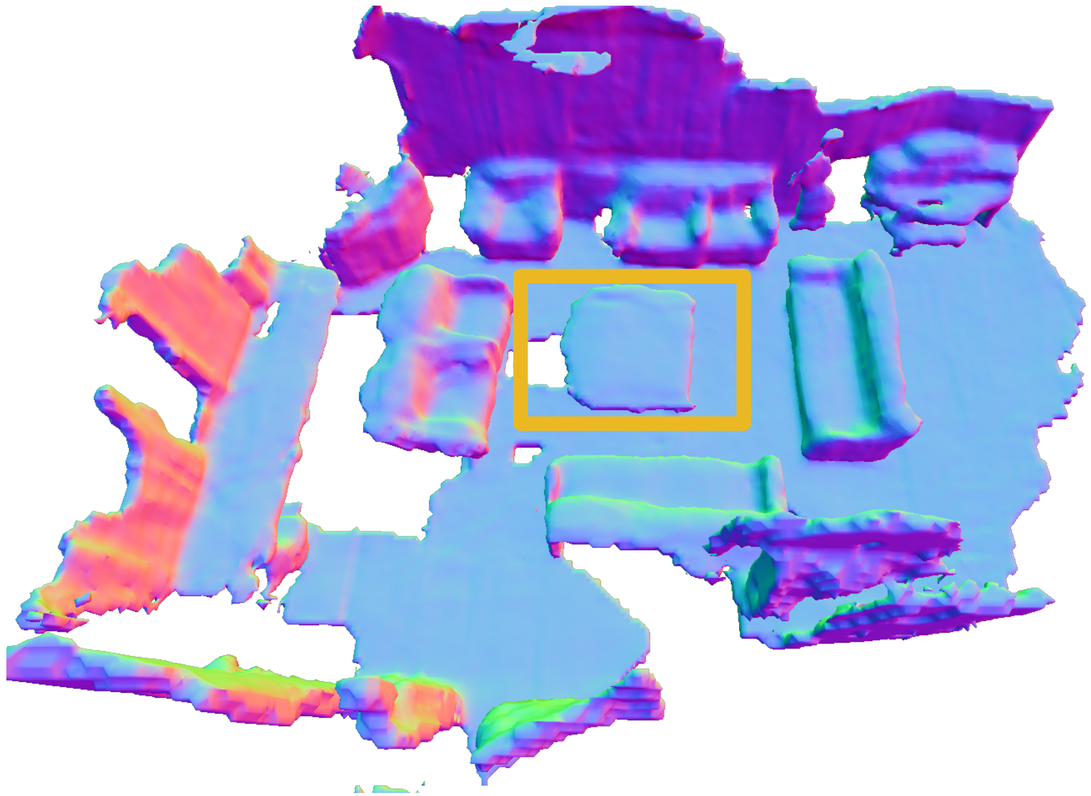
Does Number of Views Matter?



Ablation Study

Does Number of Views Matter?

Num of views = 5

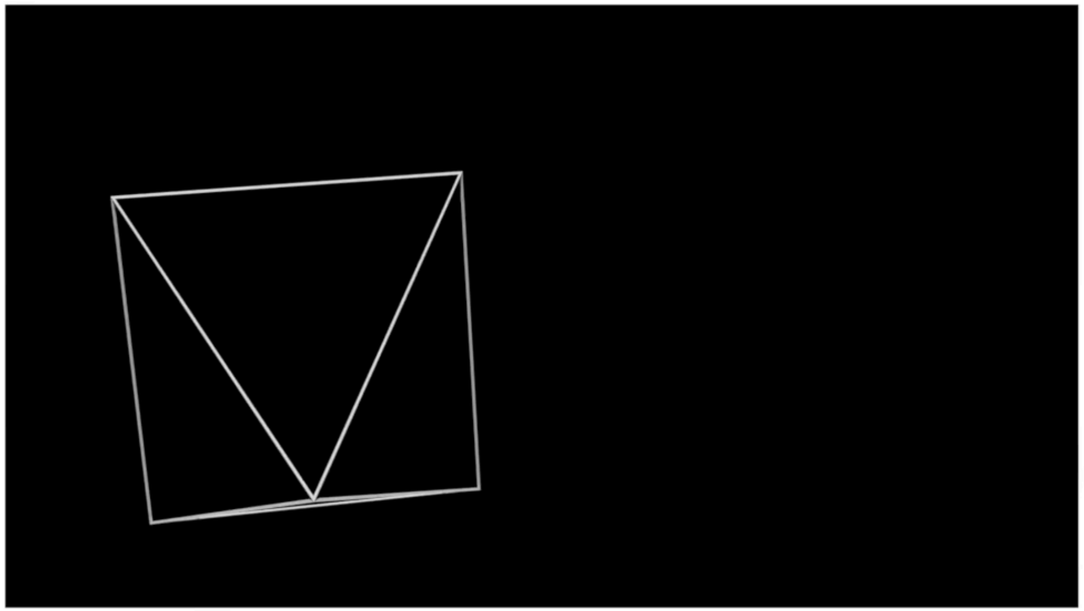


Num of views = 9

Conclusion

Project page: *https://zju3dv.github.io/neuralrecon/*

Input video



3D reconstruction