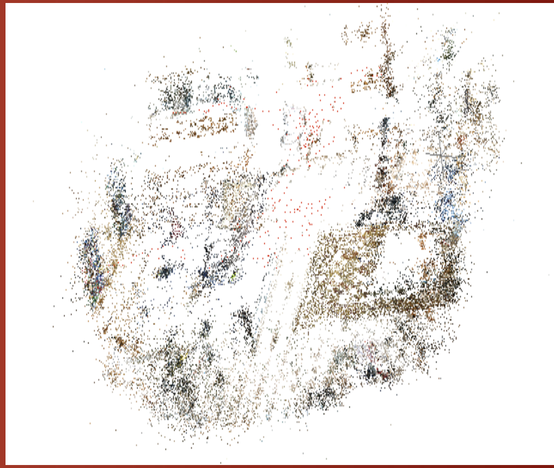
SST: Real-time End-to-end Monocular 3D Reconstruction via Sparse Spatial-Temporal Guidance

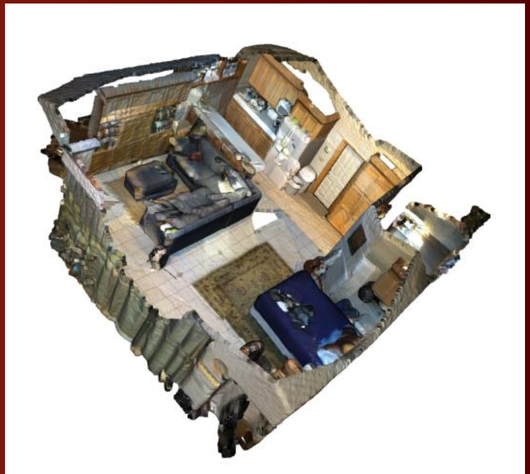
Chenyangguang Zhang, Zhiqiang Lou, Yan Di, Federico Tombari and Xiangyang Ji

IEEE International Conference on Multimedia and Expo 2023
Brisbane Convention & Exhibition Centre
10-14 July 2023

Background

- Monocular 3D scene reconstruction: predicting 3D model from consecutive frames, without distance measurements.
- Traditional visual SLAM systems: sparse reconstruction;
- Two-stage deep learning methods: spatial inconsistency.

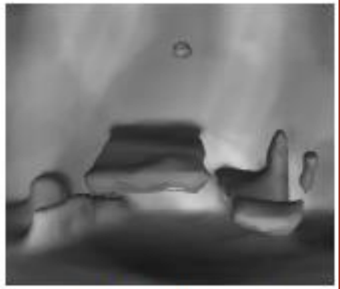


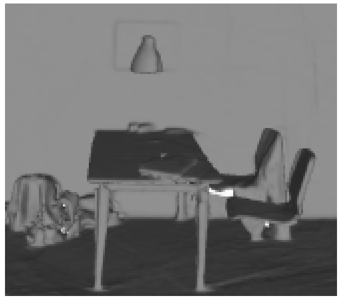


IEEE International Conference on Multimedia and Expo 2023
Brisbane Convention & Exhibition Centre
10-14 July 2023

Motivation

- End-to-end methods: directly regress TSDF volumes.
- Over-smoothed results due to insufficient supervision neglecting spatial details. (only GT TSDF with large voxel size)
- Oversimplified feature fusion ignoring temporal cues. (simple average feature fusion)


NeuralRecon


Ground Truth

Our Method

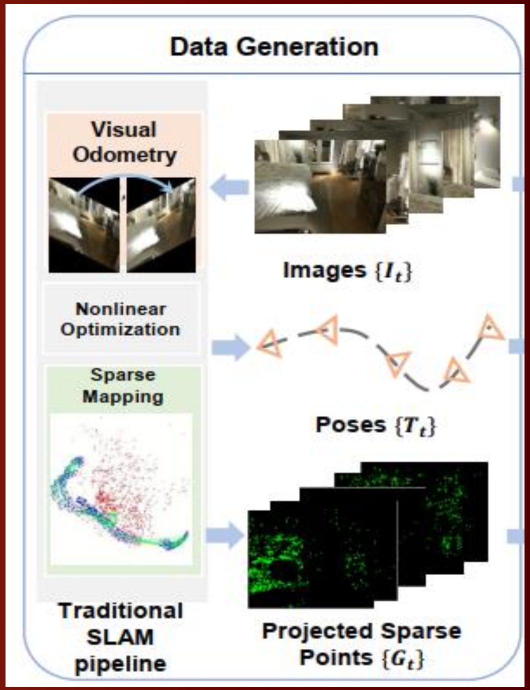
- Over-smoothed results due to insufficient supervision neglecting spatial details.
- ***-> Introducing sparse depth input – a by-product of VSLAM system for free***

- Oversimplified feature fusion ignoring temporal cues.
- ***-> Proposing sparse cross-modal attention mechanism for utilizing spatial-temporal cues***

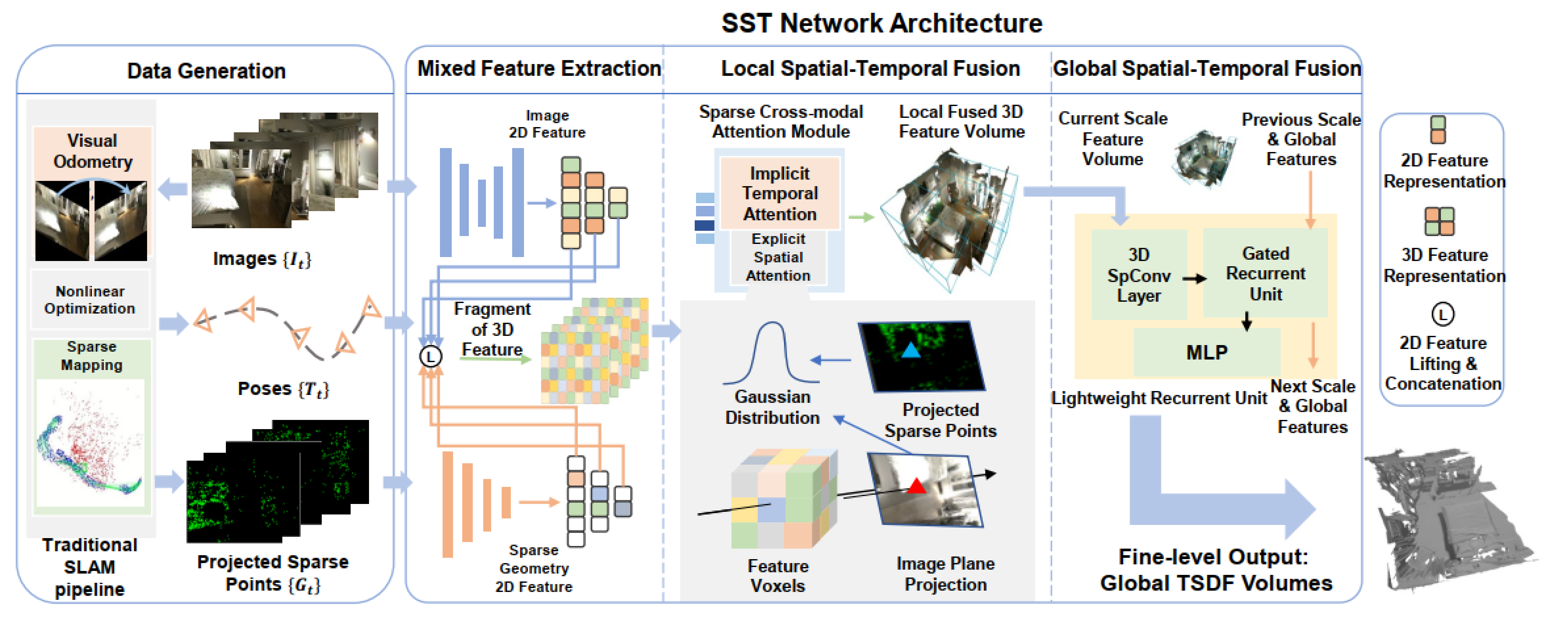
Our Method

- ***Sparse depth input – a by-product of VSLAM system for free:***
- Current end-to-end method needs a real-time off-line VSLAM system for generating camera poses.

- ***Sparse cross-modal attention mechanism:***
- Adaptive feature aggregation for color information and sparse point-based geometry priors, distilling more informative cues for accurate reconstruction.

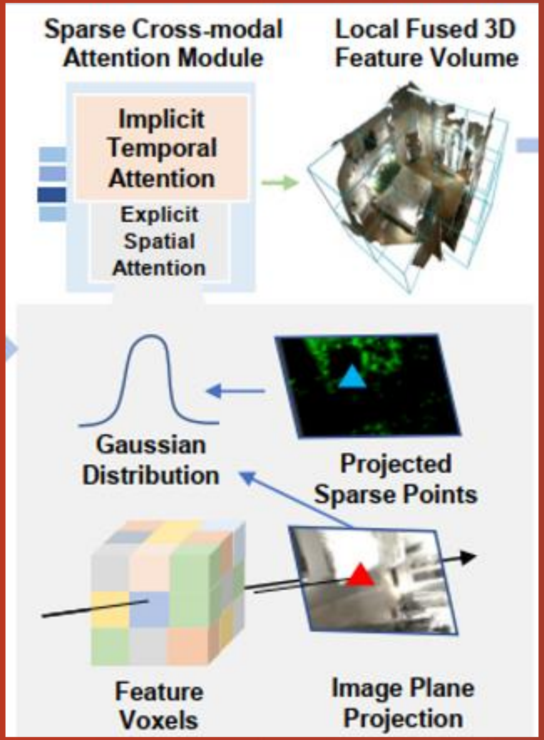


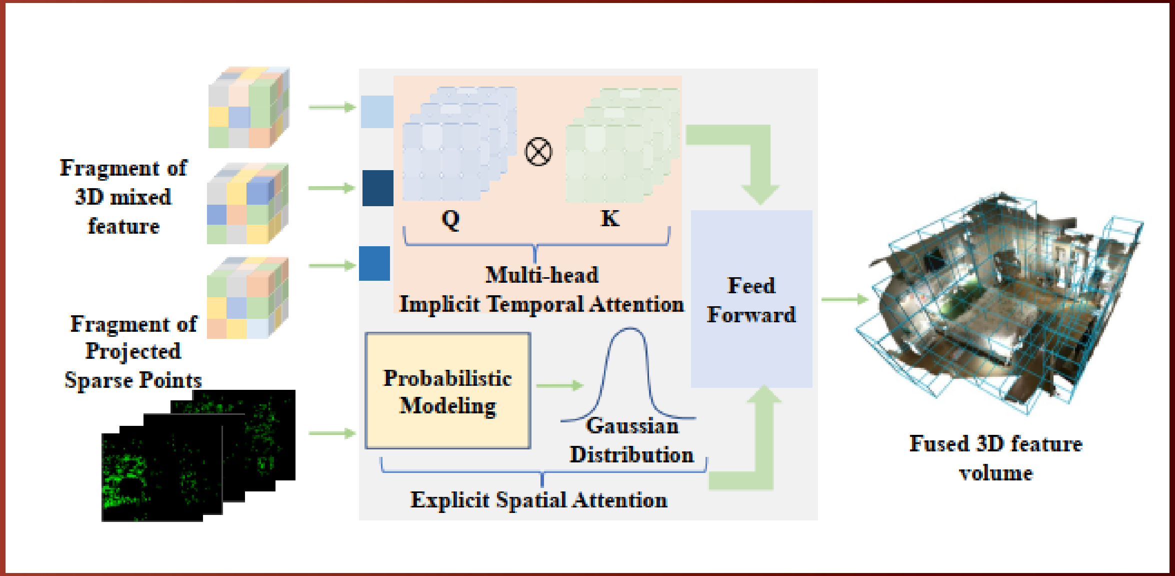
Our Method: Pipeline



Our Method: Local Spatial-Temporal Fusion

- LSTF with sparse cross-modal attention mechanism, enabling both adaptive weighted feature fusion in temporal dimension and channel-wise multi-modal feature interaction for spatial feature fusion.





Our Method: Local Spatial-Temporal Fusion

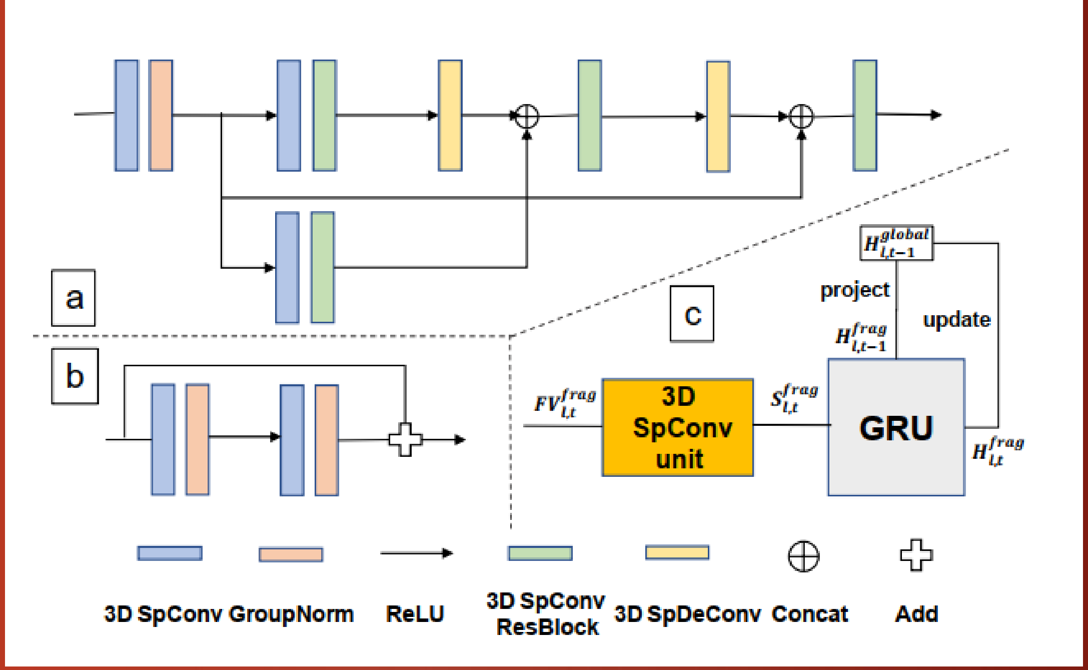
- Implicit Temporal Attention: $\omega_{l,im}$
- Explicit Spatial Attention: $\omega_{l,ex}$

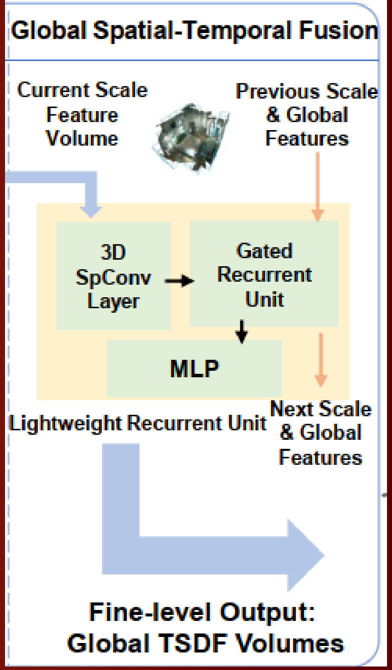
$$
\begin{aligned}
&A_{in} = [FV^{IG}_{1,v_i}, \cdots, FV^{IG}_{N,v_i}] \\
&Q = W_q A_{in},\ K = W_k A_{in},\ V = W_v A_{in} \qquad (2) \\
&\omega_{l,im} = Softmax(QK^T),\ A_{out} = \omega_{l,im}\omega_{l,ex}V
\end{aligned}
$$

$$
\omega_{l,i,t} = \begin{cases} Gauss_{\sigma_{E_t P(v_i)}}(\|SD_t(p_{v_i}) - d_{v_i}\|) & SD_t(p_{v_i}) >= 0 \\ 1 & otherwise \end{cases} \qquad (3)
$$

Our Method: Global Spatial-Temporal Fusion

- The inference time bottleneck lies in the structure of 3D sparse convolution layers.
- We significantly modify the network structure and reduce the number of parameters, so that to accelerate the whole network at a large margin.





Results

TABLE I
3D GEOMETRY METRICS ON SCANNET.

Method	Comp	Acc	Recall	Prec	F-score	FPS
MVDNet [7]	0.040	0.240	0.831	0.208	0.329	28
GPMVS [16]	**0.031**	0.879	**0.871**	0.188	0.304	27
DPSNet [8]	0.045	0.284	0.793	0.223	0.344	4
COLMAP [12]	0.069	0.135	0.634	0.505	0.558	0.4
NeuralRecon [3]	0.138	0.053	0.472	0.687	0.559	47
Ours	0.124	**0.053**	0.505	**0.695**	**0.584**	**59**
TransFusion [1]	0.082	0.055	0.600	**0.728**	**0.655**	7
Atlas [2]	0.076	0.071	**0.605**	0.675	0.636	4
NeuralRecon [3]	0.075	0.051	0.556	0.706	0.621	47
Ours	**0.071**	**0.050**	0.584	0.714	0.643	**59**

TABLE II
2D DEPTH METRICS ON SCANNET.

Method	Abs.Rel.	Abs.Diff.	Sq.Rel.	RMSE	$\delta < 1.25$
MVDNet [7]	0.098	0.191	0.061	0.293	89.6
GPMVS [16]	0.130	0.239	0.339	0.472	90.6
DPSNet [8]	0.087	0.158	0.035	0.232	92.5
COLMAP [12]	0.137	0.264	0.138	0.502	83.4
Atlas [2]	0.065	0.123	0.045	0.251	93.6
NeuralRecon [3]	0.065	0.099	0.034	0.197	93.7
Ours	**0.060**	**0.092**	**0.034**	**0.185**	**94.0**

TABLE III
3D METRICS ON 7-SCENES.

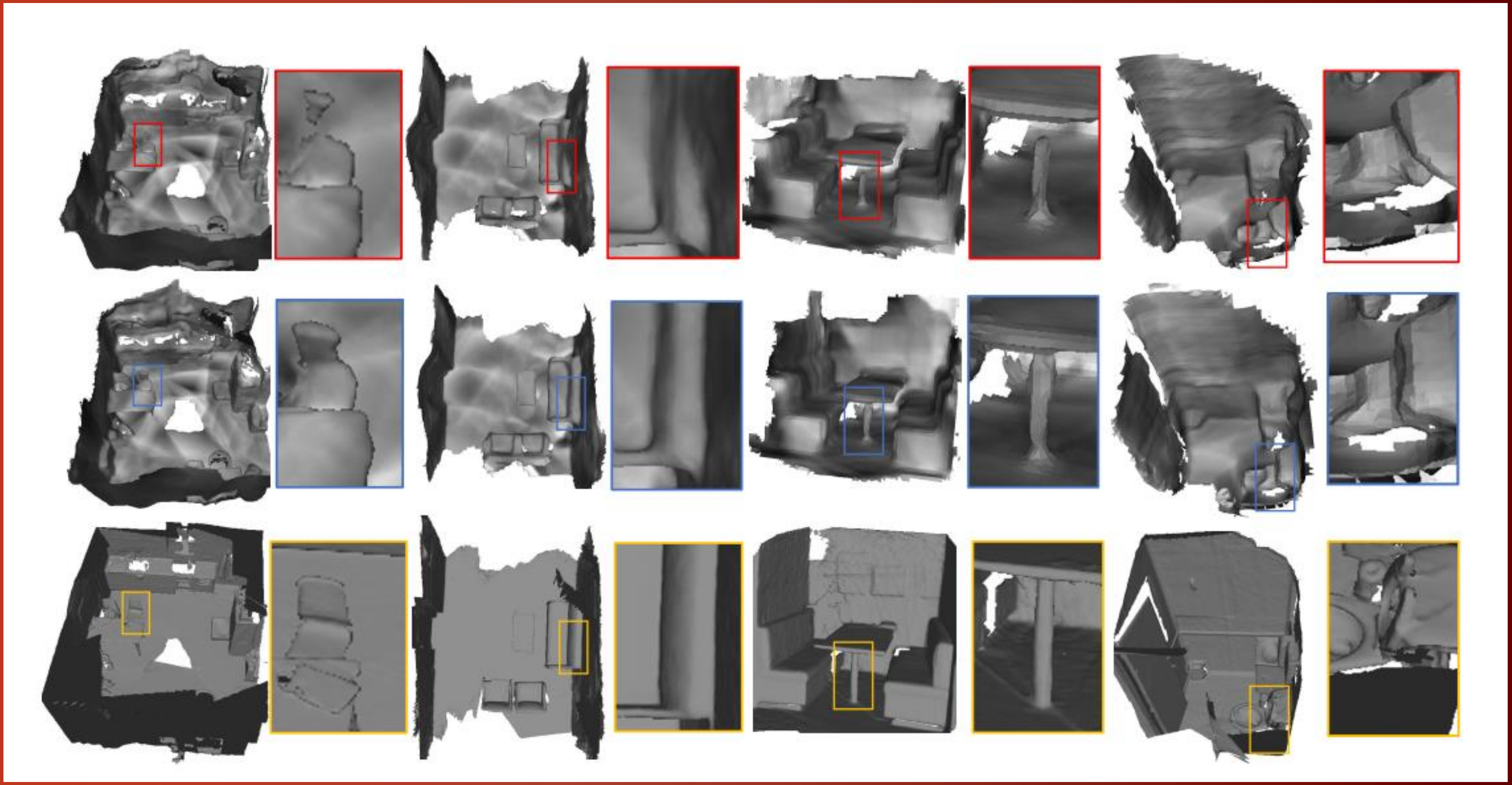
Method	Comp	Acc	Recall	Prec	F-score	FPS
NeuralRecon [3]	0.228	**0.100**	0.228	0.389	0.282	47
Ours	**0.225**	0.104	**0.242**	**0.392**	**0.298**	**59**

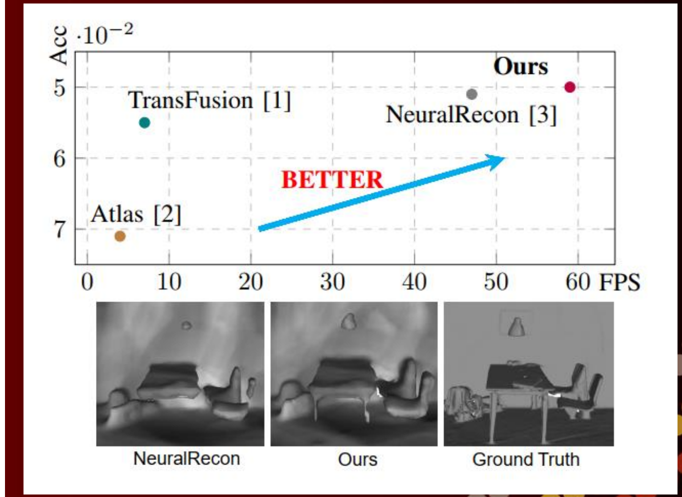
TABLE IV
LSTF ,FE AND GSTF ARCHITECTURE ABLATIONS ON SCANNET UNDER 3D METRICS ALONG WITH THE TOP BLOCK OF TAB. I.

	SCAM	Geometry Priors	LWRU	Recall	Prec	F-score	FPS
a	×	×	×	0.472	0.687	0.559	47
b	✓	×	×	0.494	0.690	0.574	41
c	✓	×	✓	0.495	0.695	0.576	**62**
d	✓	✓	×	0.496	0.695	0.579	38
e	✓	✓	✓	**0.505**	**0.695**	**0.584**	59

IEEE International Conference on Multimedia and Expo 2023
Brisbane Convention & Exhibition Centre
10-14 July 2023

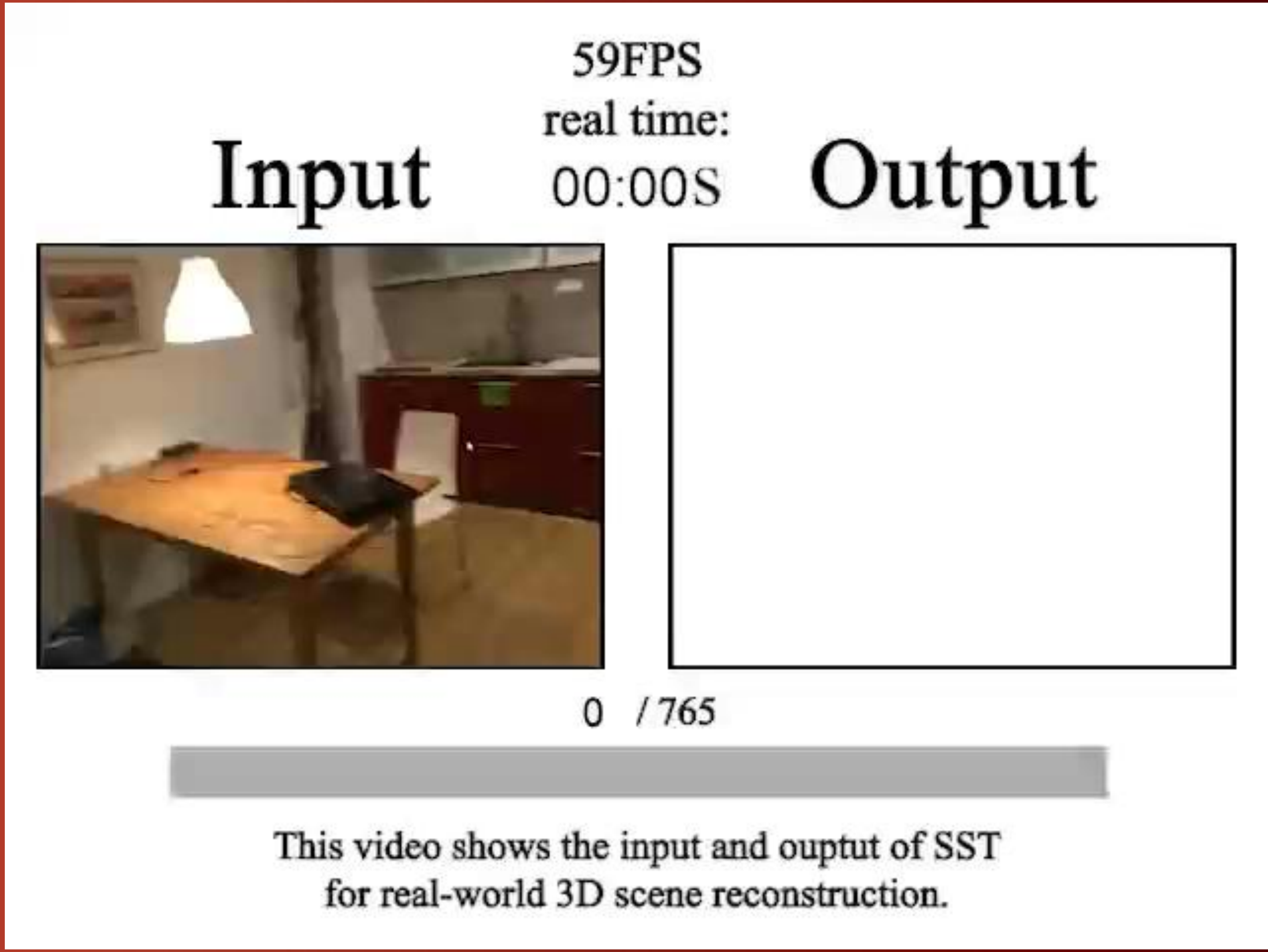
Results





Acc $\cdot 10^{-2}$
Ours
TransFusion [1]
NeuralRecon [3]
BETTER
Atlas [2]
5
6
7
0 10 20 30 40 50 60 FPS
NeuralRecon
Ours
Ground Truth

Demo Video


59FPS
real time:
00:00S
Input
Output
0 / 765
This video shows the input and ouptut of SST
for real-world 3D scene reconstruction.

IEEE International Conference on Multimedia and Expo 2023

Brisbane Convention & Exhibition Centre

10-14 July 2023

IEEE Advancing Technology for Humanity · CAS · IEEE ComSoc · IEEE Computer Society · IEEE Signal Processing Society

Thanks for your attention!

www.2023.ieeeicme.org

zcyg22@mails.tsinghua.edu.cn