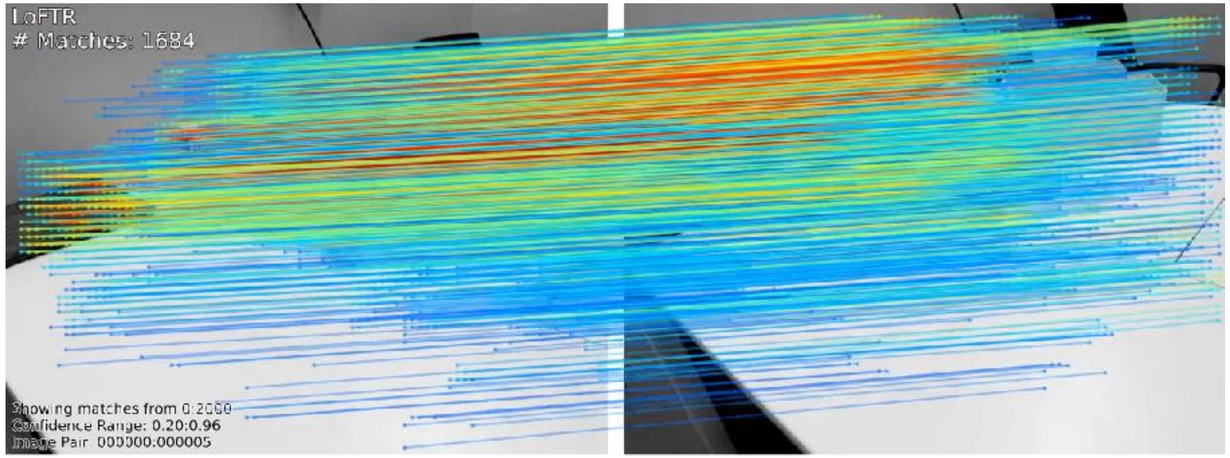
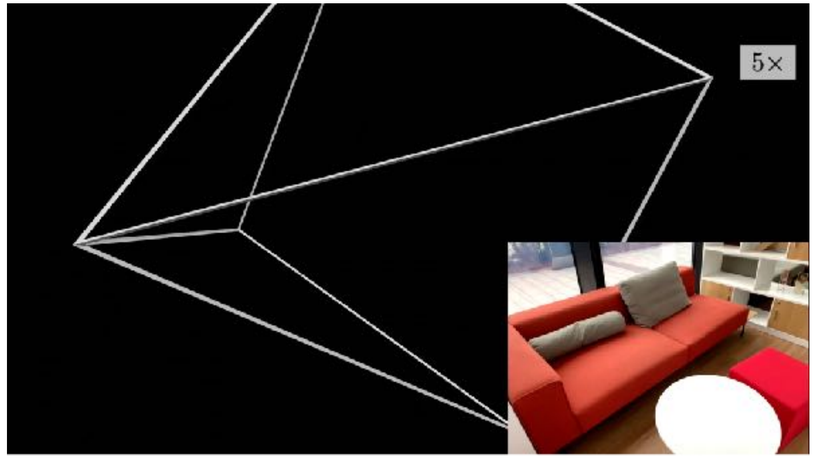
Learning-based 3D Reconstruction and Localization: What Should be Learned?

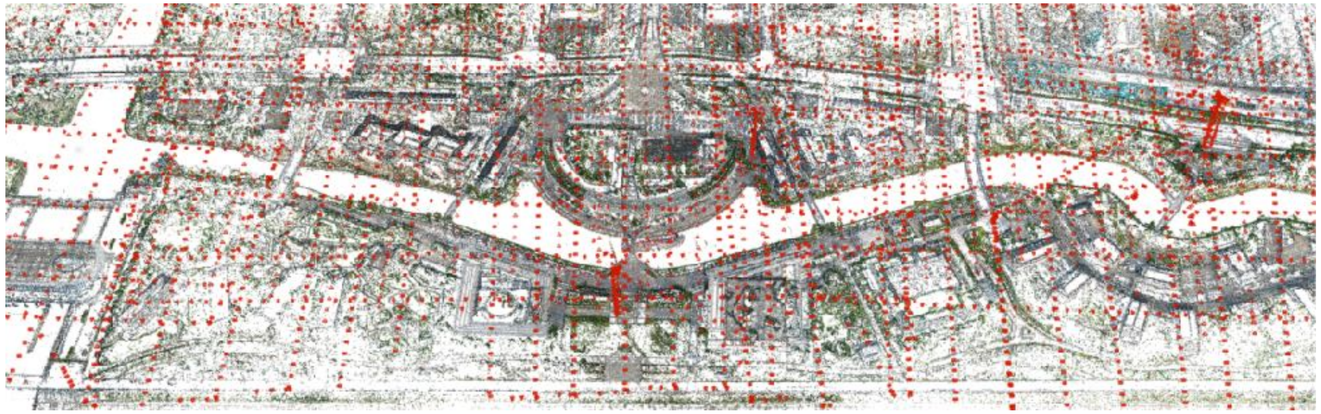
Jiaming Sun
SenseTime & ZJU-3DV

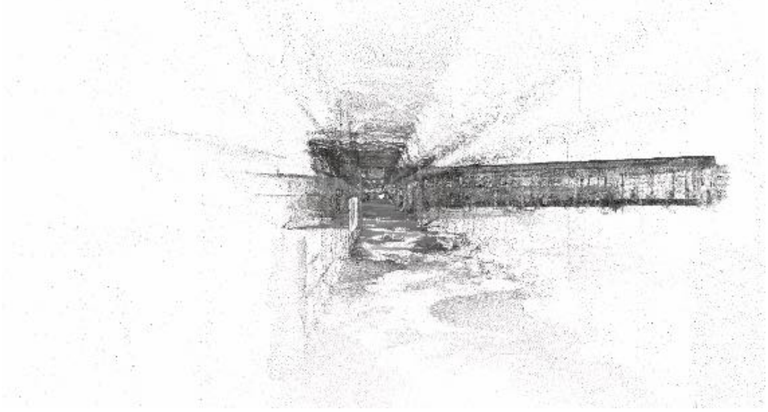




Applications at SenseTime

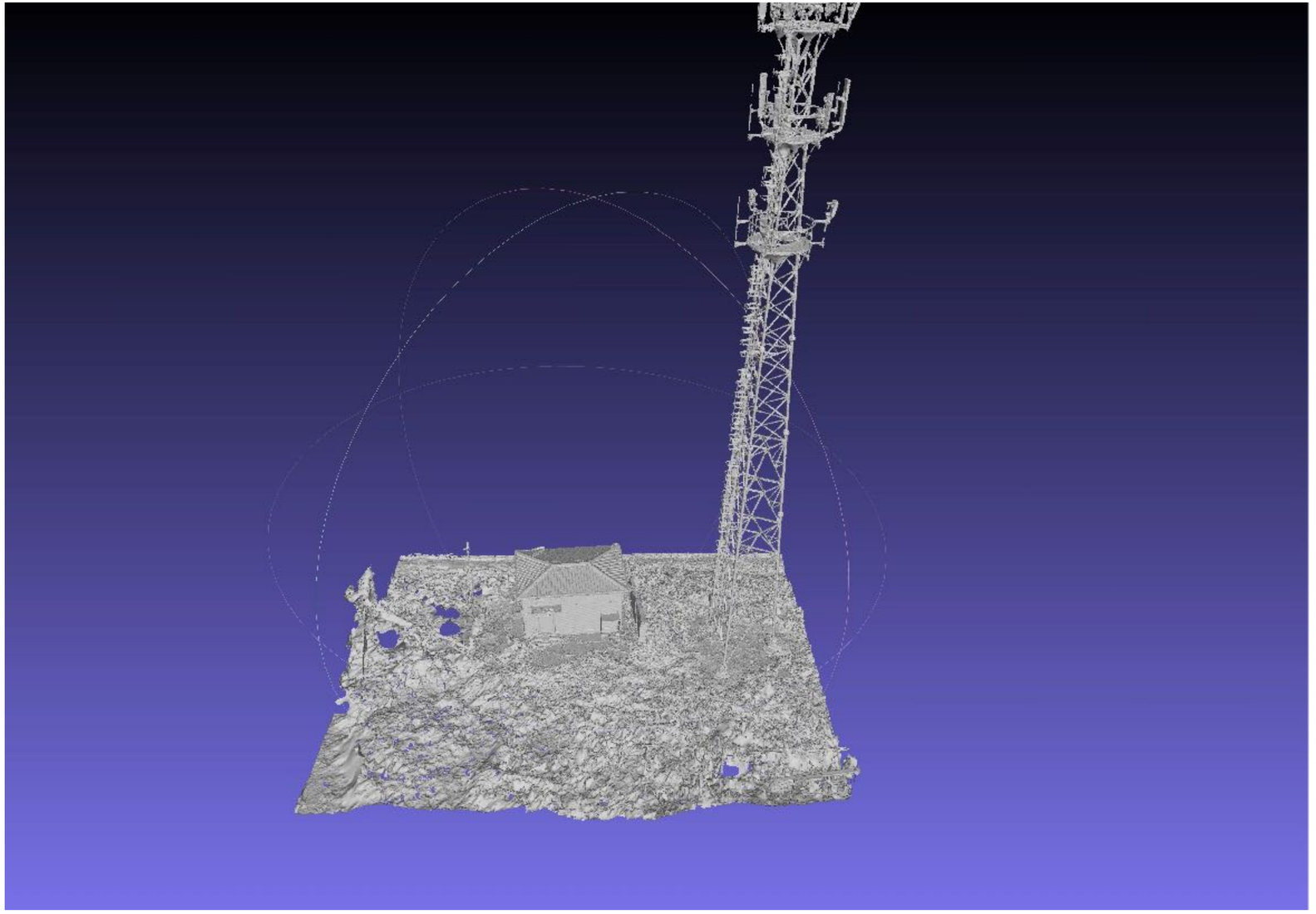
Localization

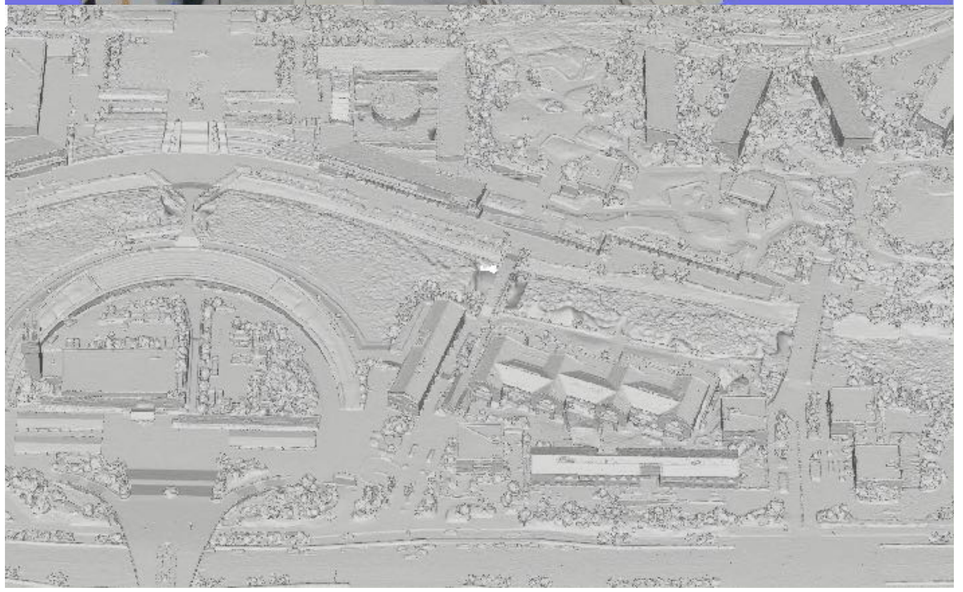




Applications at SenseTime

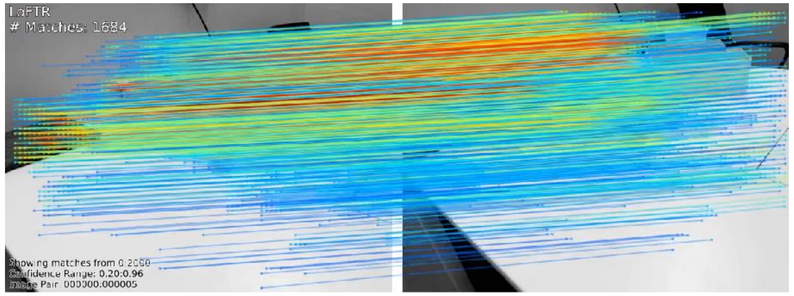
3D reconstruction





Today's Topic

Visual localization

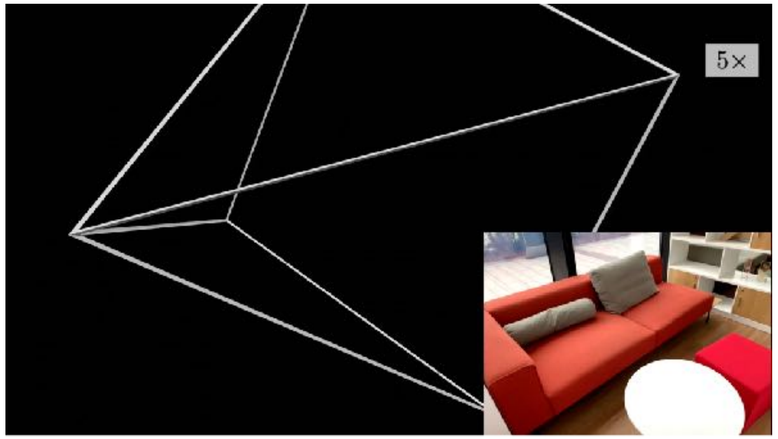


LoFTR: Detector-Free Local Feature Matching with Transformers

Jiaming Sun* Zehong Shen* Yu'ang Wang* Hujun Bao Xiaowei Zhou

CVPR 2021

3D Reconstruction



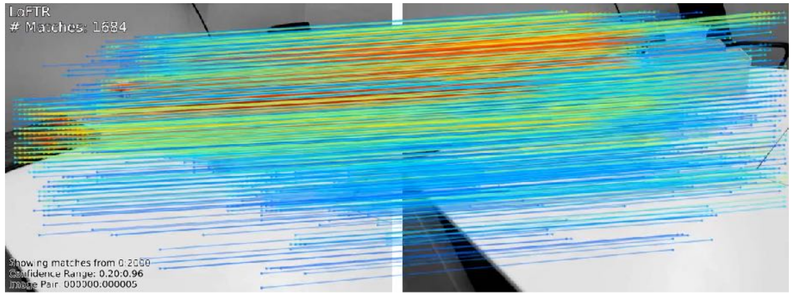
NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video

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

CVPR 2021 (Oral)

Today's Topic

Visual localization

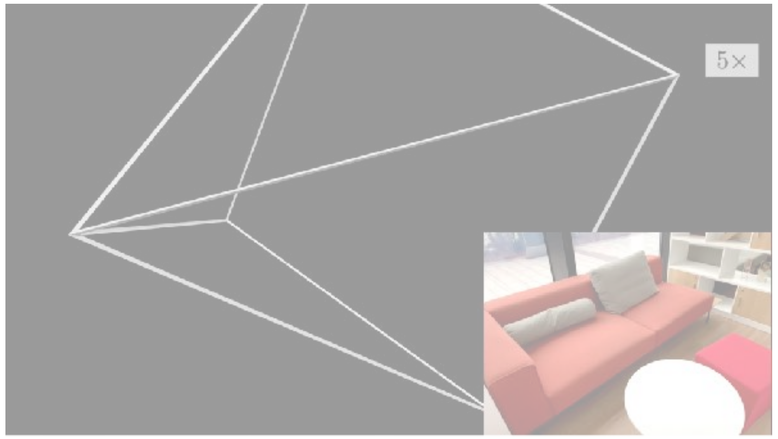


LoFTR: Detector-Free Local Feature Matching with Transformers

Jiaming Sun* Zehong Shen* Yu'ang Wang* Hujun Bao Xiaowei Zhou

CVPR 2021

3D Reconstruction

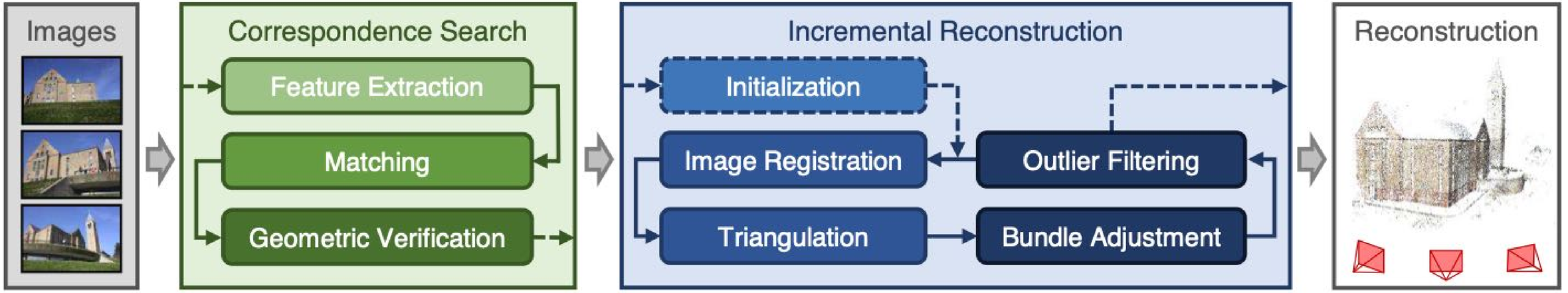


NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video

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

CVPR 2021 (Oral)

Structure-from-Motion (SfM)



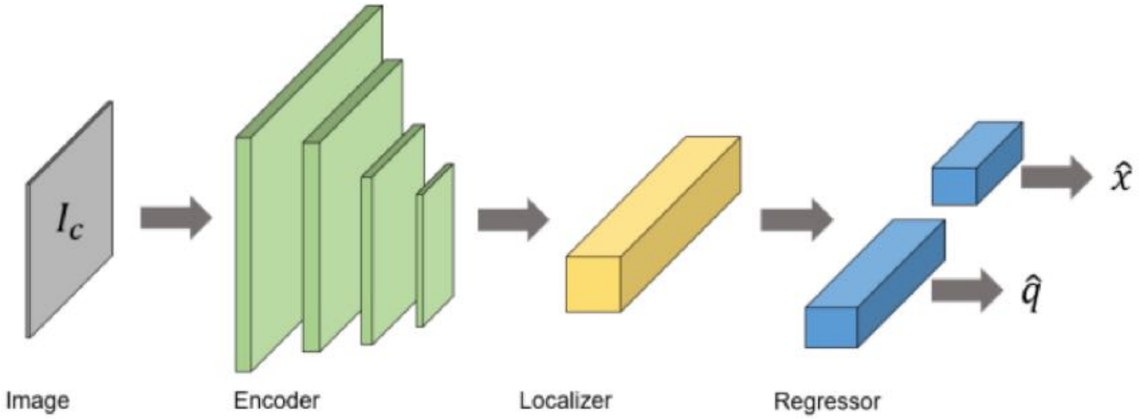
Credit: COLMAP

Direct Camera Pose Regression

Absolute pose estimation

PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization

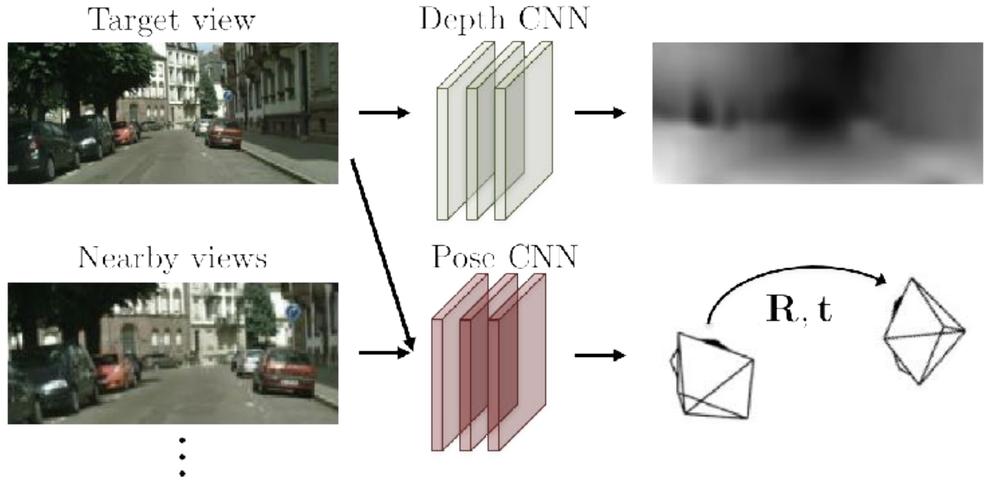
Alex Kendall
Matthew Grimes
University of Cambridge
Roberto Cipolla



Relative pose estimation

Unsupervised Learning of Depth and Ego-Motion from Video

Tinghui Zhou*
UC Berkeley
Matthew Brown
Google
Noah Snavely
Google
David G. Lowe
Google



(b) Testing: single-view depth and multi-view pose estimation.

Direct Camera Pose Regression

Understanding the Limitations of CNN-based Absolute Camera Pose Regression

Torsten Sattler[1] Qunjie Zhou[2] Marc Pollefeys[3,4] Laura Leal-Taixé[2]
[1]Chalmers University of Technology [2]TU Munich [3]ETH Zürich [4]Microsoft

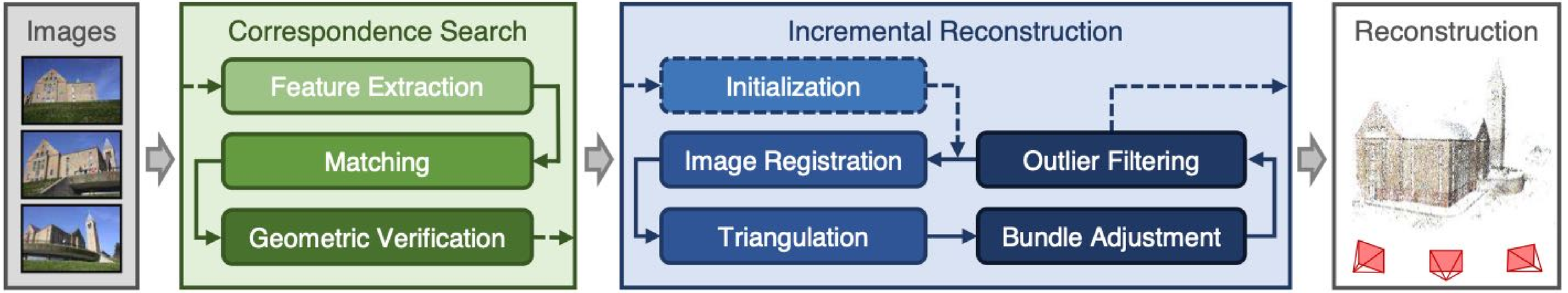
of APR techniques. Based on this model, we show that APR approaches are more closely related to approximate pose estimation via image retrieval (Sec. 5) than to accurate pose estimation via 3D geometry (Sec. 4). **ii)** Using

To Learn or Not to Learn: Visual Localization from Essential Matrices

Qunjie Zhou[1], Torsten Sattler[2], Marc Pollefeys[3,4], Laura Leal-Taixé[1]

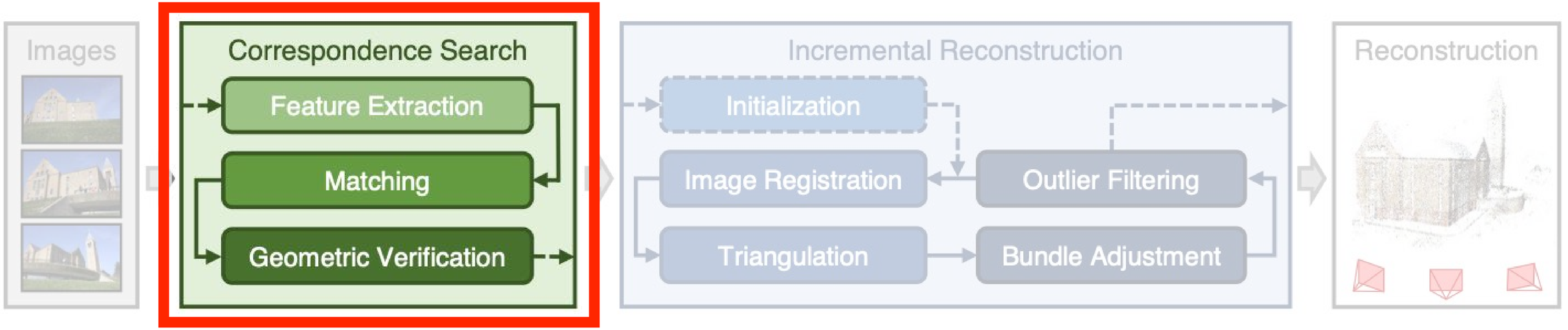
ing their results, we have shown that the purely data-driven approach does not generalize well and have identified the reason for this failure as the relative pose regression layers.

Revisiting Structure-from-Motion (SfM)



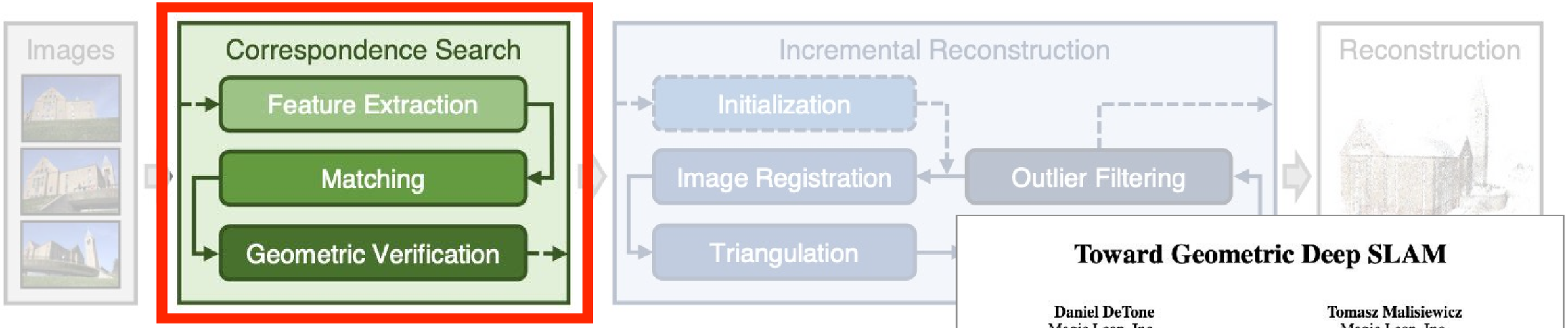
Credit: COLMAP

Learning to Find Correspondences



Credit: COLMAP

Learning to Find Correspondences



MatchNet: Unifying Feature and Metric Learning for Patch-Based Matching

Xufeng Han[†] Thomas Leung[‡] Yangqing Jia[‡] Rahul Sukthankar[‡] Alexander C. Berg[†]
[†]University of North Carolina at Chapel Hill [‡]Google Research

Toward Geometric Deep SLAM

Daniel DeTone
Magic Leap, Inc.
Sunnyvale, CA
ddetone@magicleap.com

Tomasz Malisiewicz
Magic Leap, Inc.
Sunnyvale, CA
tmalisiewicz@magicleap.com

Andrew Rabinovich
Magic Leap, Inc.
Sunnyvale, CA
arabinovich@magicleap.com

LIFT: Learned Invariant Feature Transform

Kwang Moo Yi*,[1], Eduard Trulls*,[1], Vincent Lepetit[2], Pascal Fua[1]

SuperPoint: Self-Supervised Interest Point Detection and Description

Daniel DeTone
Magic Leap
Sunnyvale, CA
ddetone@magicleap.com

Tomasz Malisiewicz
Magic Leap
Sunnyvale, CA
tmalisiewicz@magicleap.com

Andrew Rabinovich
Magic Leap
Sunnyvale, CA
arabinovich@magicleap.com

Recently: D2-Net, R2D2, GIFT, DISK, SuperGlue...

Credit: COLMAP

Local Feature Applications @ SenseTime

Benchmark测试

- 自研特征检测提取模块SVCNN相比superpoint定位成功率提升~10个点
- 整体耗时10ms

B5地下停车场					
	# query	成功率	<0.5m	<1.0m	<2.0m
V1（SIFT）	13284	71.3339%	65.53%	68.4131%	69.9262%
V2（SuperPoint）	13284	65.0407%	58.6646%	61.8714%	63.5125%
V3（SVCNN）	13284	**74.39%**	**66.2376%**	**70.1144%**	**72.3050%**

国博F1					
	# query	成功率	<0.5m	<1.0m	<2.0m
V1（SIFT）	13212	30.5101%	16.3185%	20.7236%	24.3264%
V2（SuperPoint）	13212	40.0318%	24.7427%	29.7381%	32.8716%
V3（SVCNN）	13212	**51.2413%**	**29.1023%**	**36.1792%**	**41.0157%**

Applications

商汤SenseMARS助力成都IFS打造全国首个全场景城市综合体AR导航

此次在成都IFS覆盖区域多达46万平方米，综合定位成功率高达99%，定位精度达“厘米”级别，单次时长达“毫秒”级别，且不会出现偏移、闪烁等情况。

Local Feature Applications @ SenseTime

Benchmark测试

- 自研特征检测提取模块SVCNN相比superpoint定位成功率提升~10个点
- 整体耗时10ms

	# query	成功率	<0.5m	<1.0m	<2.0m
V1（SIFT）	13284	71.3339%	65.53%	69.[illegible]%	[illegible]%
V2（SuperPoint）	13284	65.0407%	58.6646%	61.8714%	63.5125%
V3（SVCNN）	13284	**74.39%**	**66.2376%**	**70.1144%**	**72.3050%**

国博F1

	# query	成功率	<0.5m	<1.0m	<2.0m
V1（SIFT）	13212	30.5101%	16.3185%	20.7236%	24.3264%
V2（SuperPoint）	13212	40.0318%	24.7427%	29.7381%	32.8716%
V3（SVCNN）	13212	**51.2413%**	**29.1023%**	**36.1792%**	**41.0157%**

OK, this is all pretty cool and exciting.
What's next?

Applications

商汤SenseMARS助力成都IFS打造全国首个全场景城市综合体AR导航

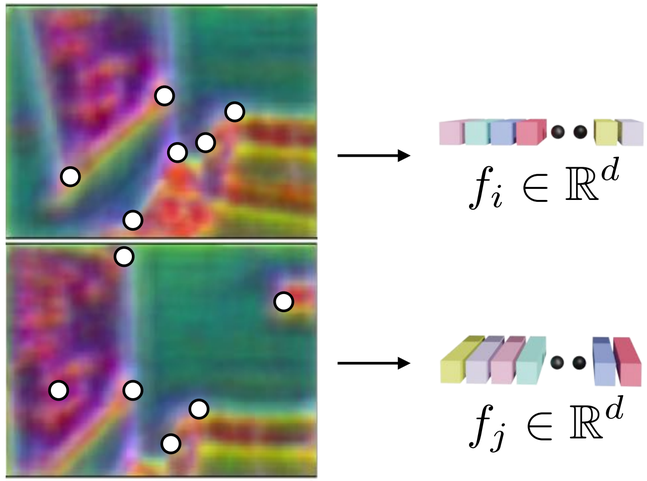
此次在成都IFS覆盖区域多达46万平方米，综合定位成功率高达99%，定位精度达“厘米”级别，单次时长达“毫秒”级别，且不会出现偏移、闪烁等情况。

Introduction to LoFTR

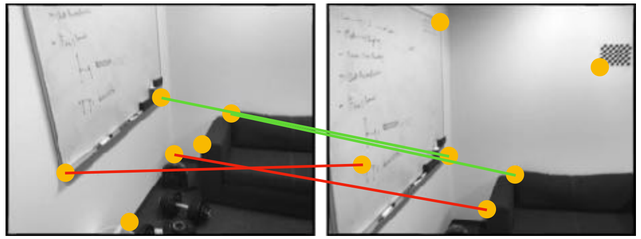
The standard pipeline for local feature matching

Input images

Feature detection



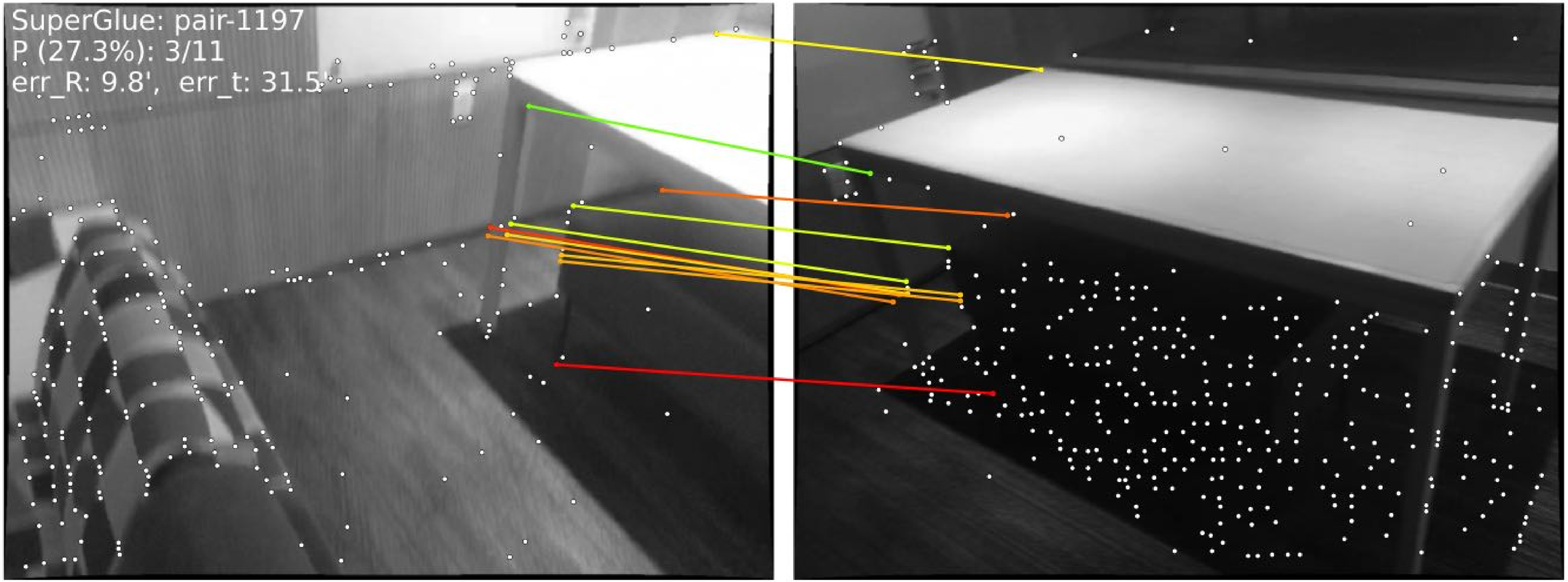
Feature description



Matching with
Mutual-Nearest-Neighbor (MNN)

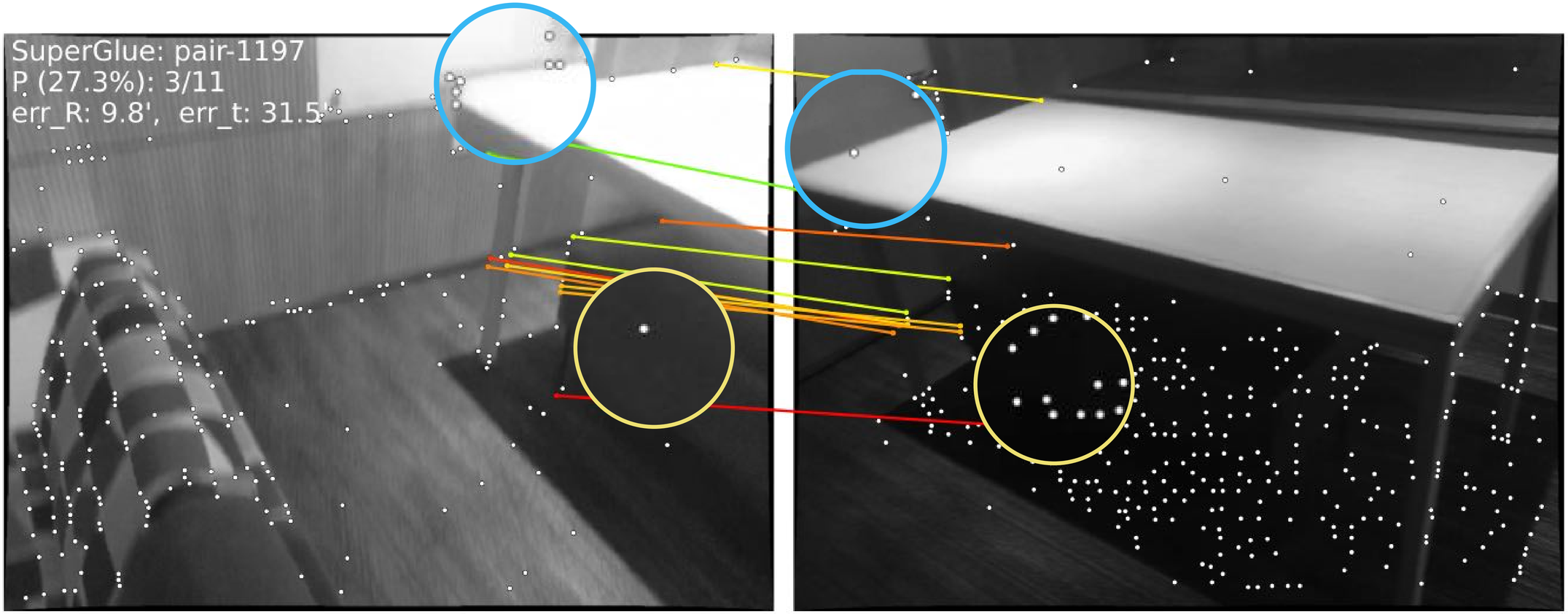
Introduction to LoFTR

Challenge 1: Detector repeatability on two (or more) images



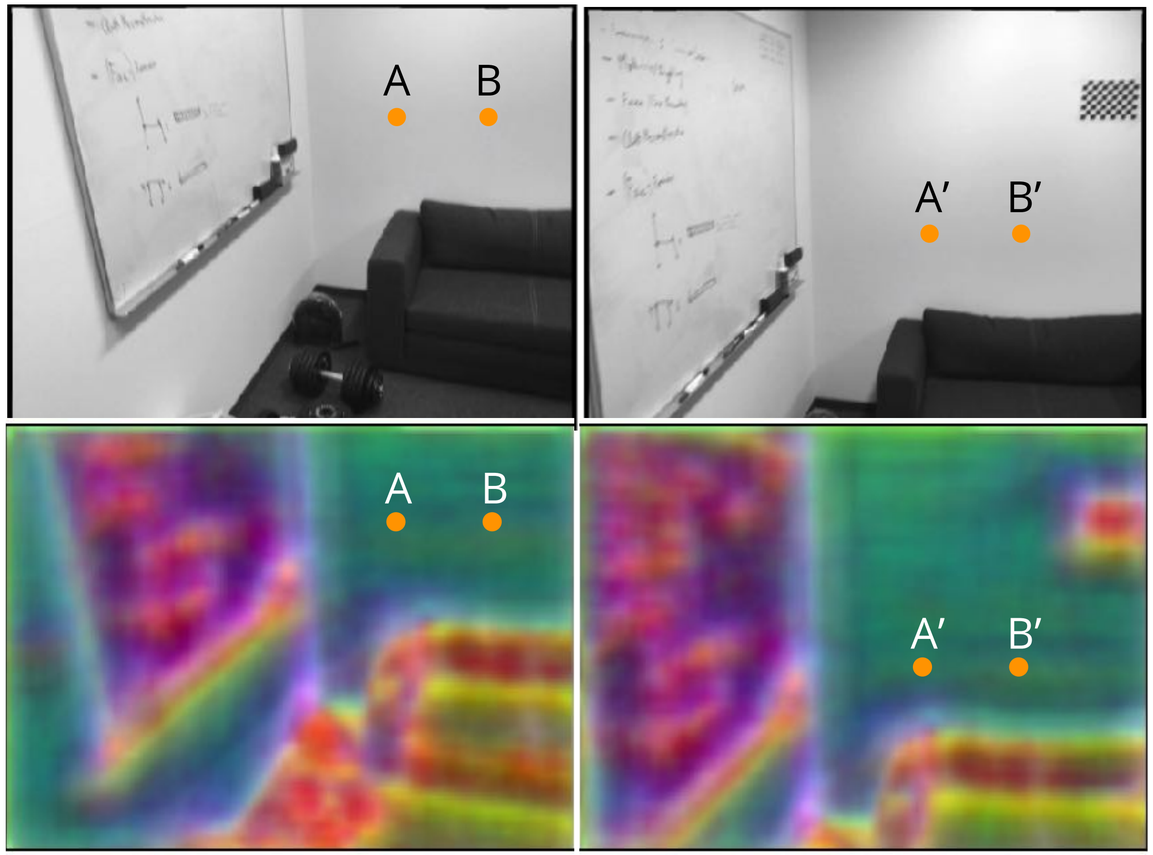
Introduction to LoFTR

Challenge 1: Detector repeatability on two (or more) images

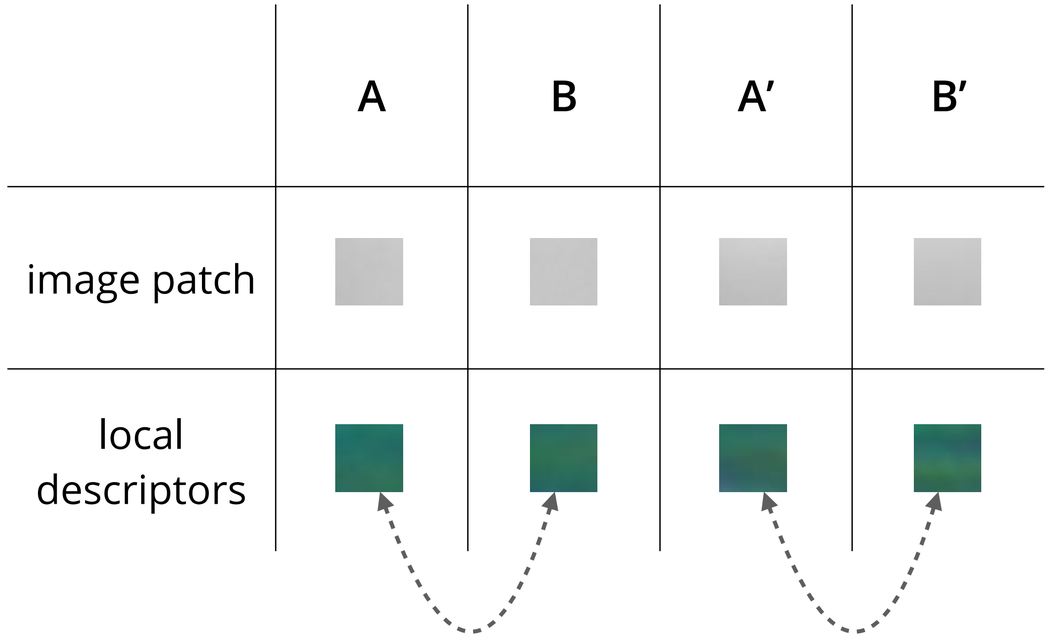


Introduction to LoFTR

Challenge 2: Local descriptors are not position-dependent



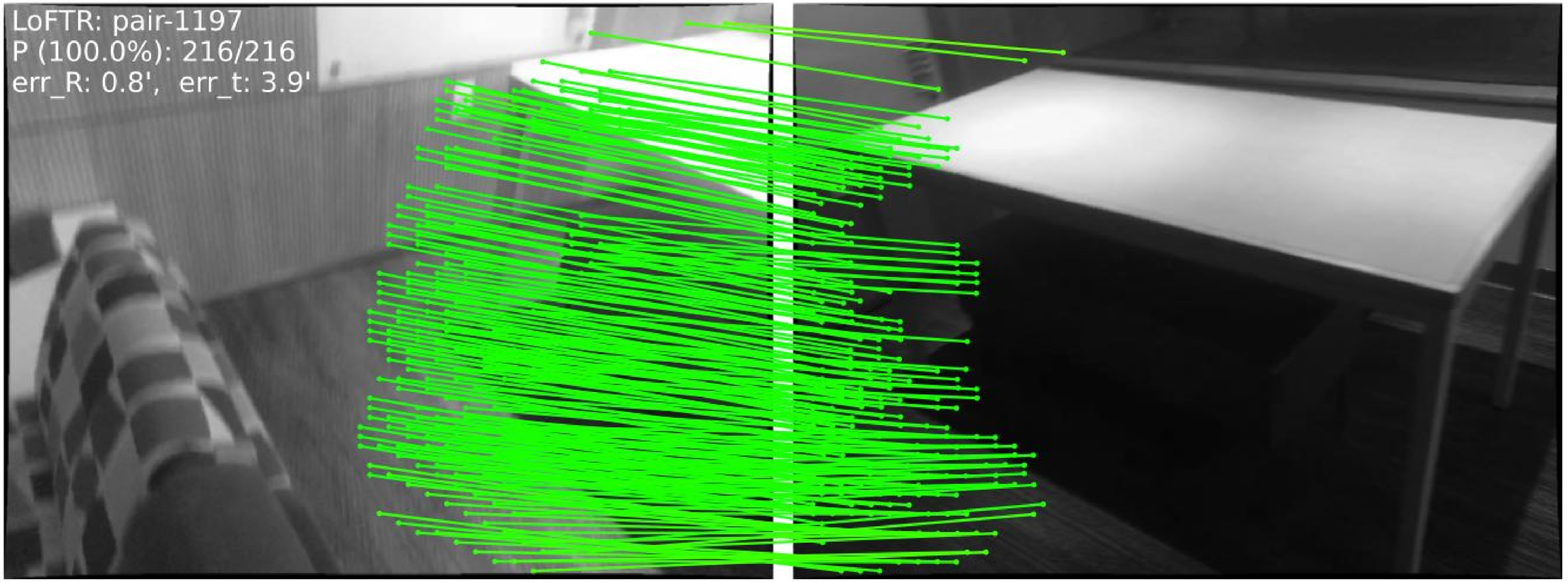
Local features visualized with PCA



Expected to be different (position-dependent)

Introduction to LoFTR

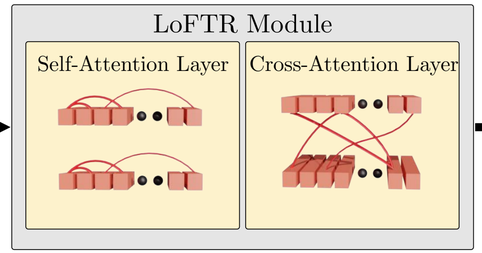
Our solution: LoFTR



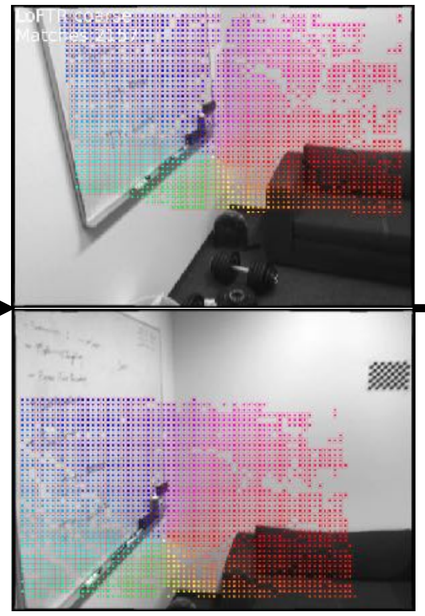
LoFTR

Pipeline overview

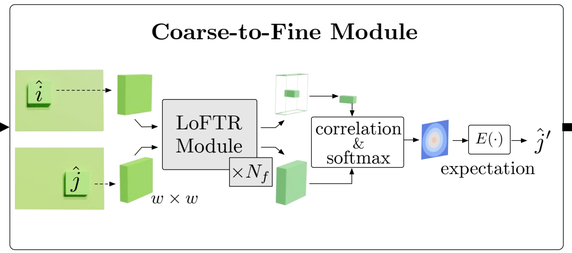
Input image



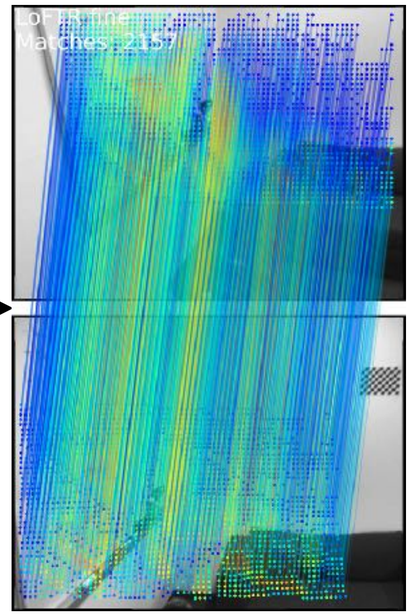
LoFTR coarse-matching



Dense coarse matches



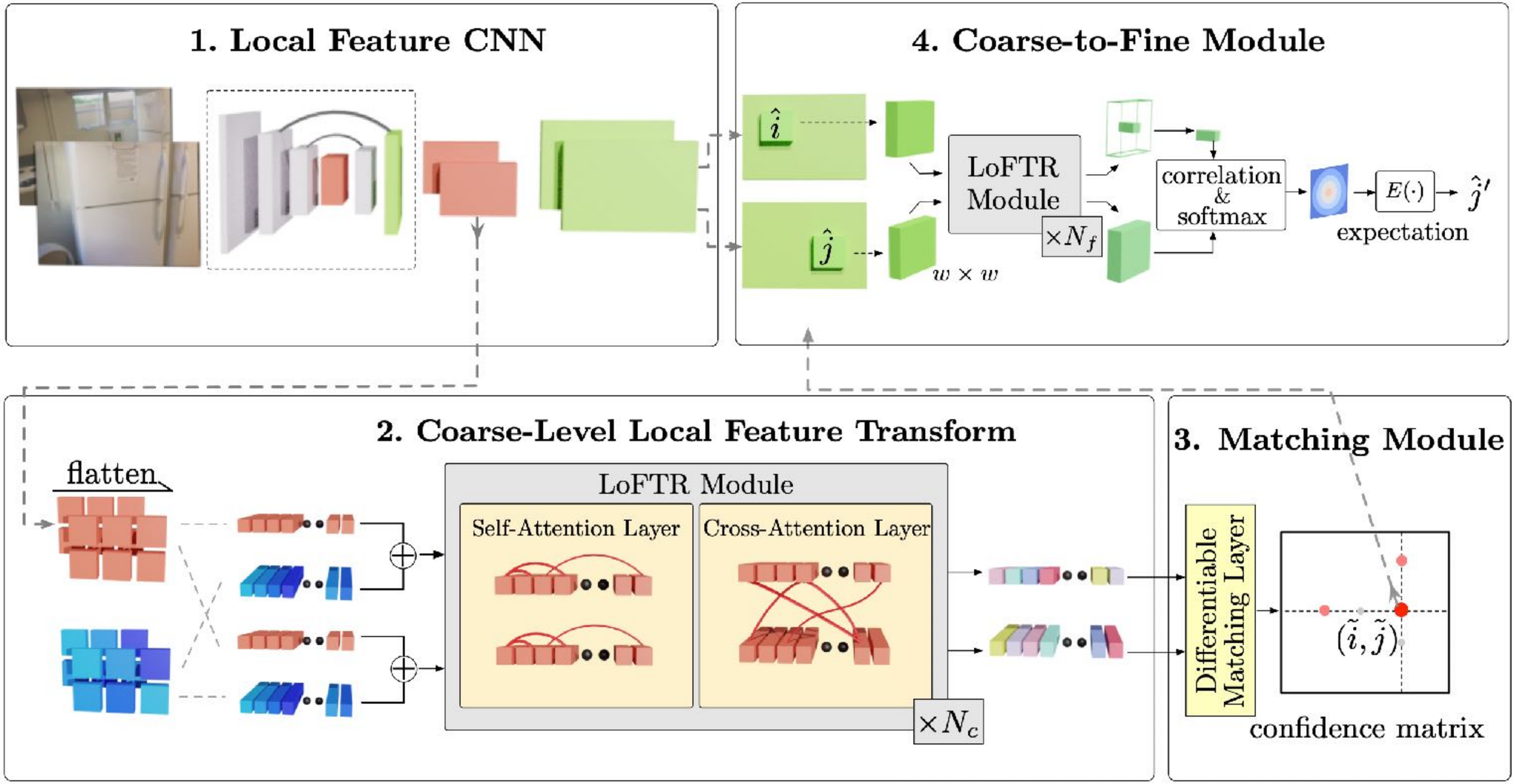
LoFTR fine-matching



Final matches

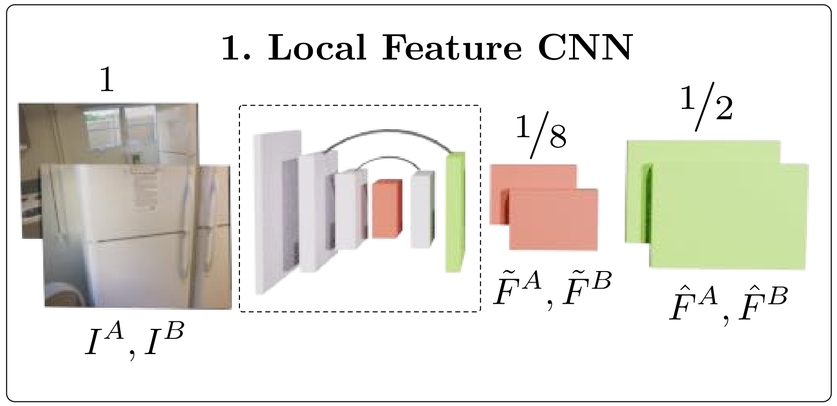
LoFTR

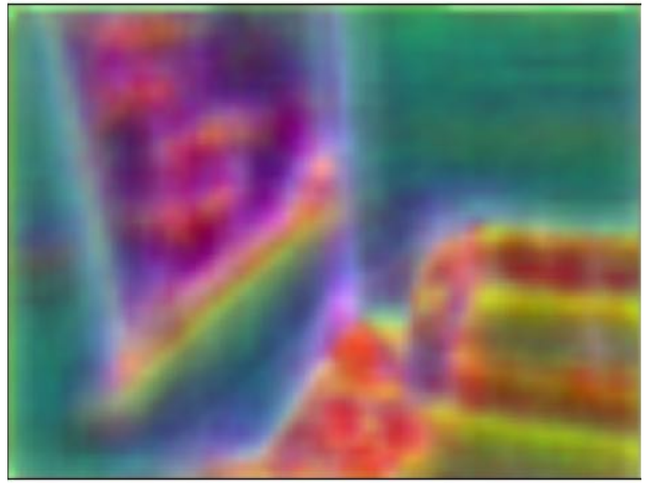
Architecture

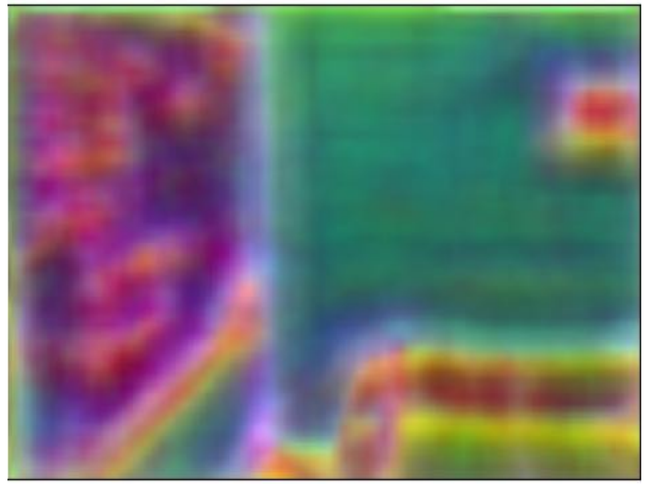


LoFTR

Local feature CNN

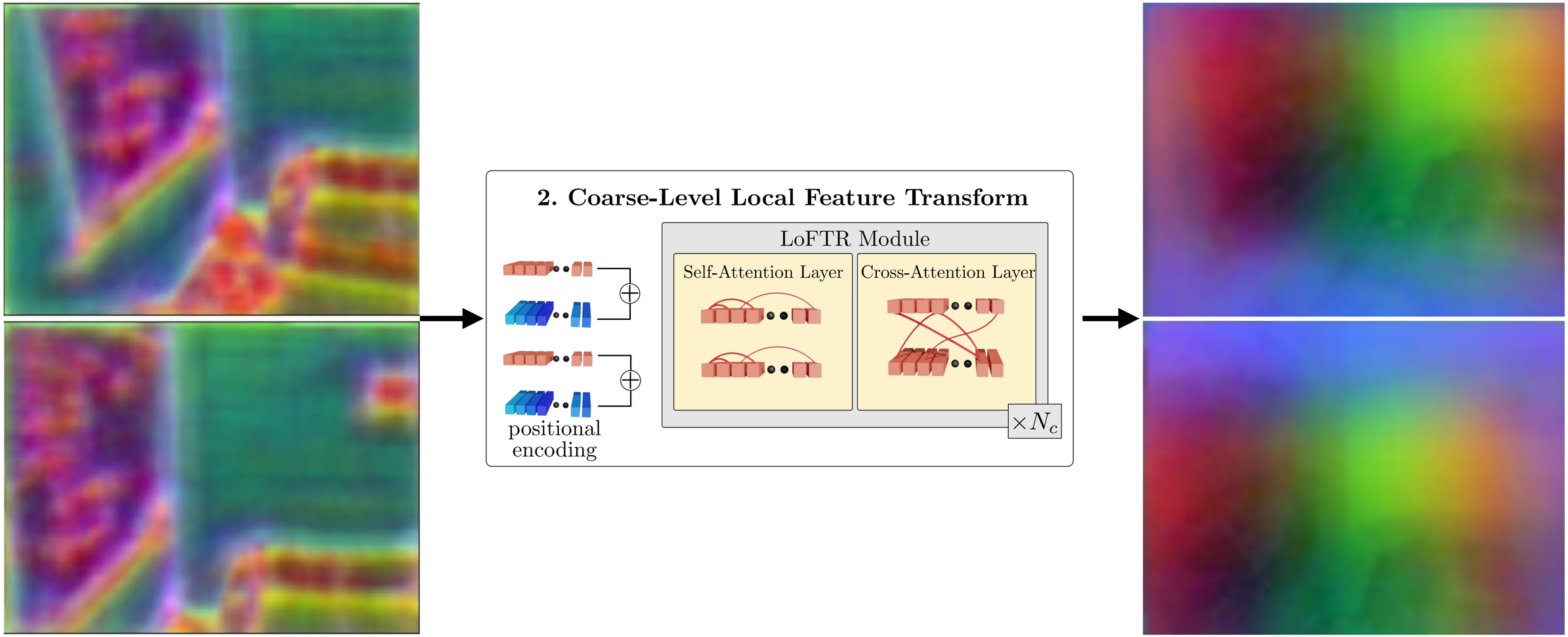






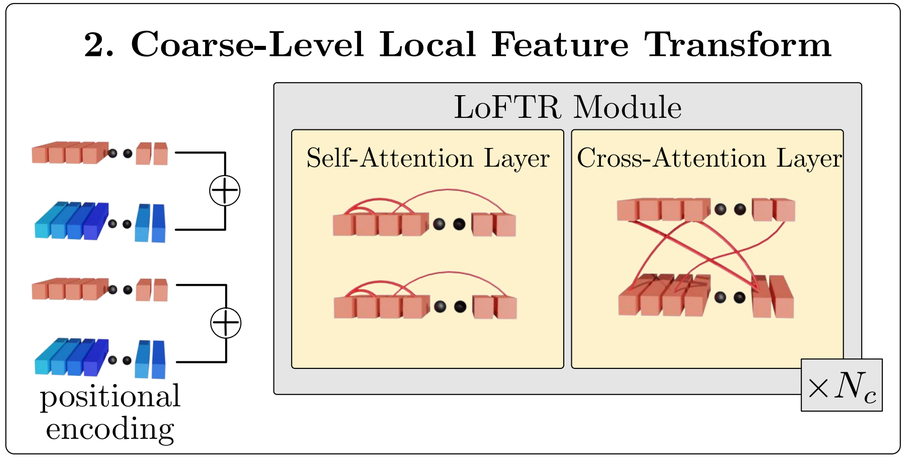
LoFTR

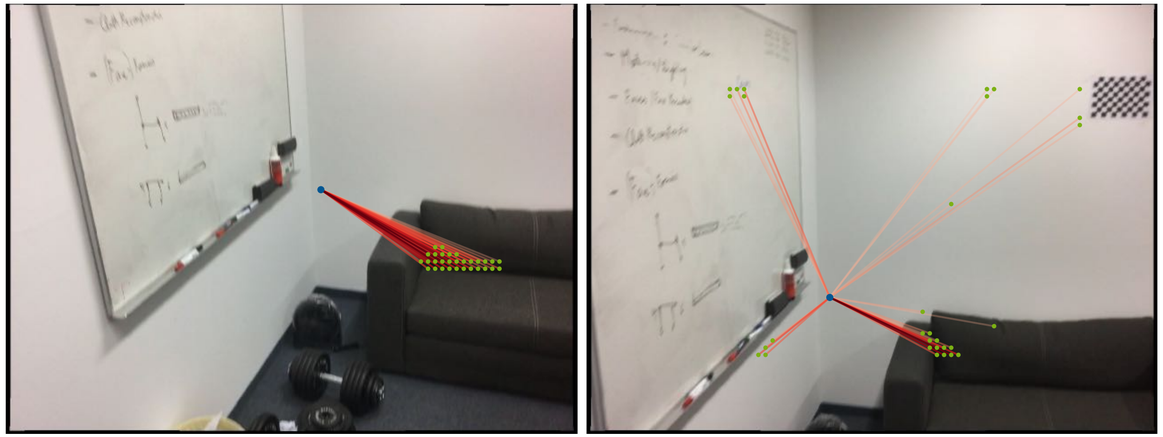
Coarse-level LoFTR module



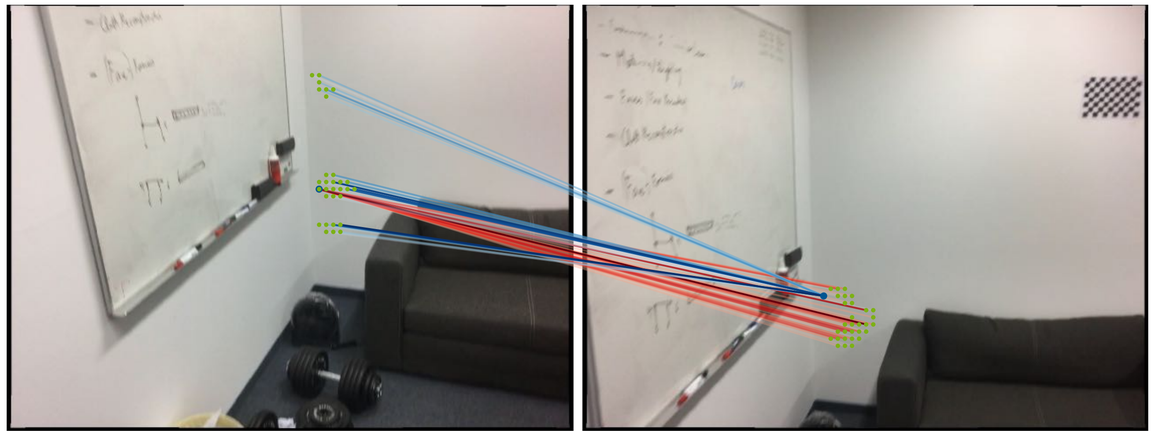
LoFTR

Coarse-level LoFTR module





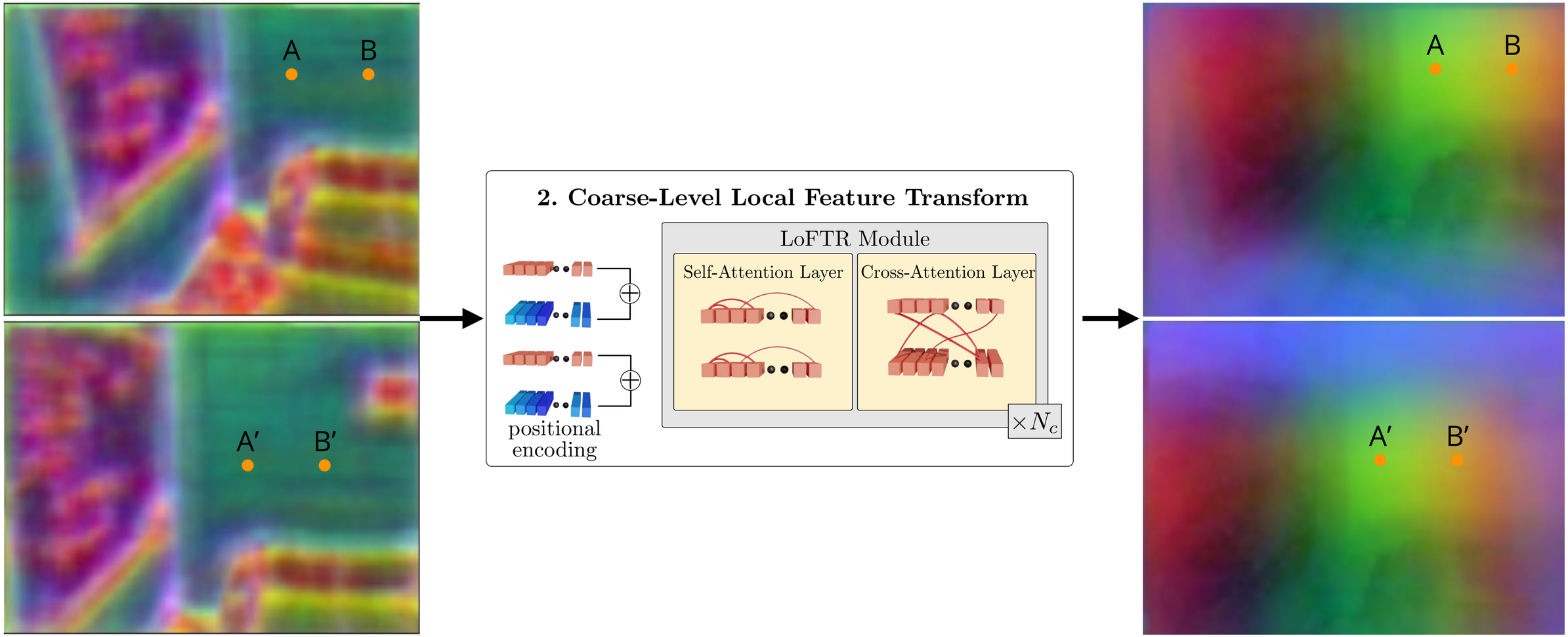
Attention weight visualization of **self-attention**



Attention weight visualization of **cross-attention**

LoFTR

Coarse-level LoFTR module

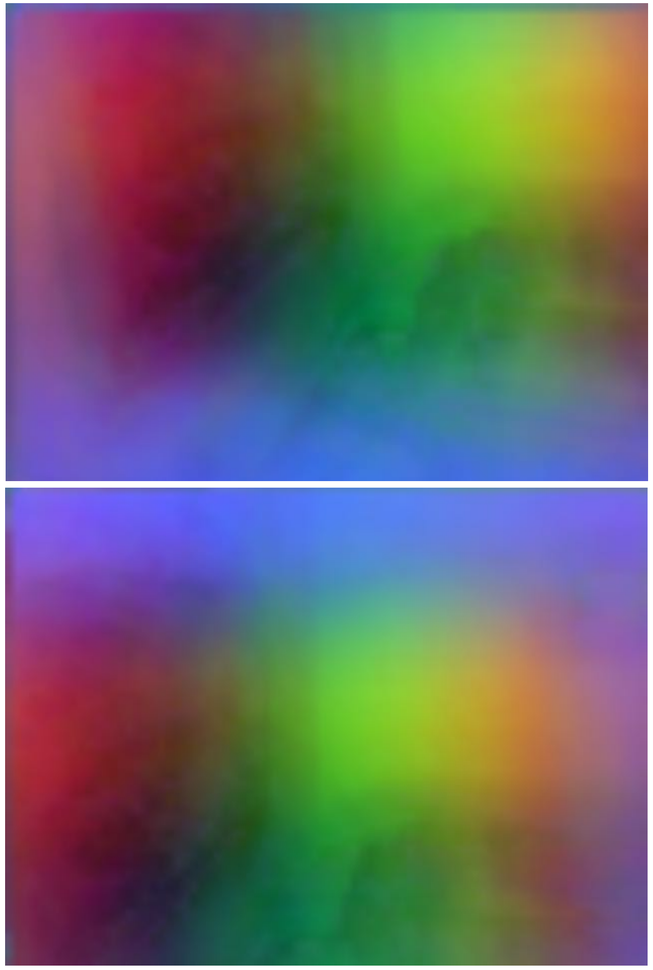


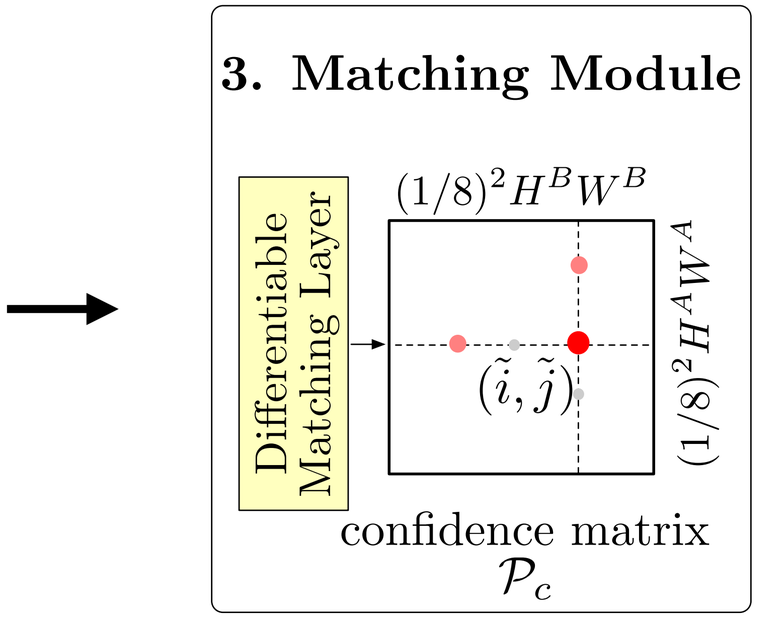
Local features

Transformed features

LoFTR

Differentiable matching layer



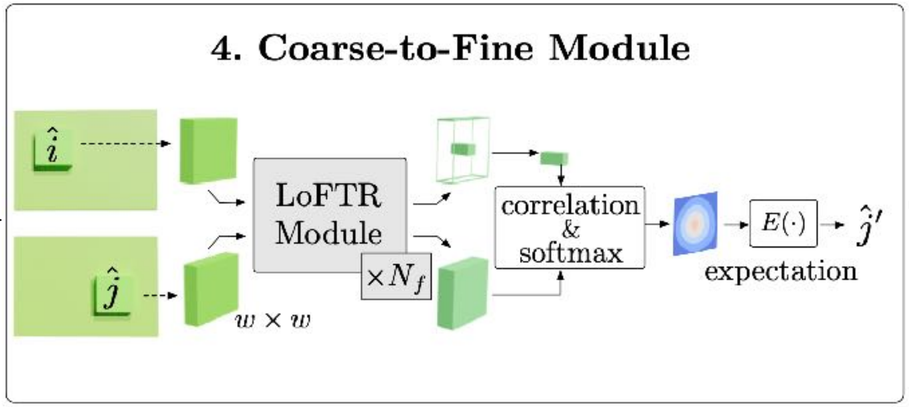


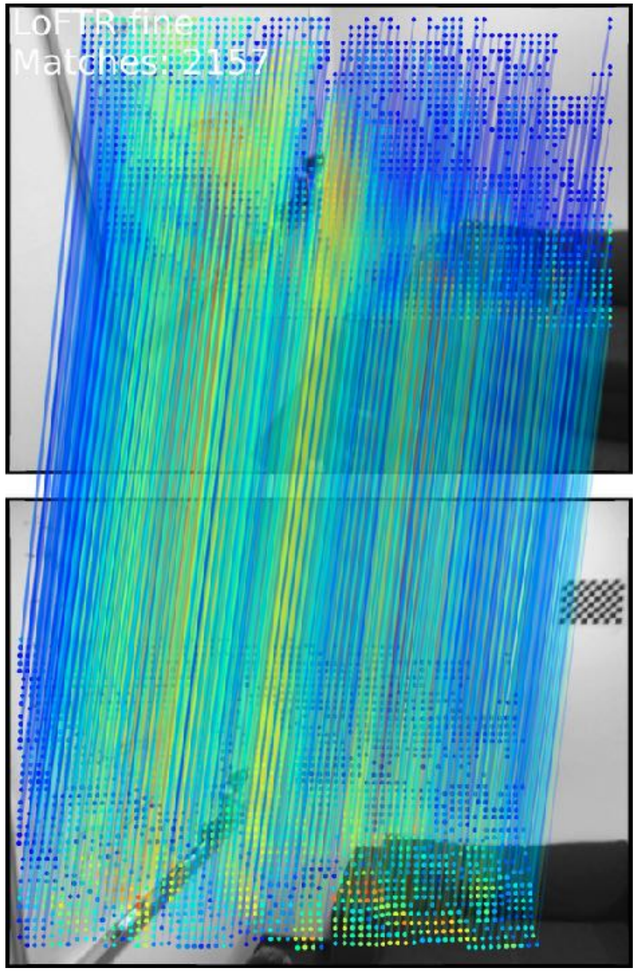
$$\mathcal{P}_c(i,j) = \text{softmax}\left(\mathcal{S}\left(i,\cdot\right)\right)_j \cdot \text{softmax}\left(\mathcal{S}\left(\cdot,j\right)\right)_i$$

Dual-softmax

LoFTR

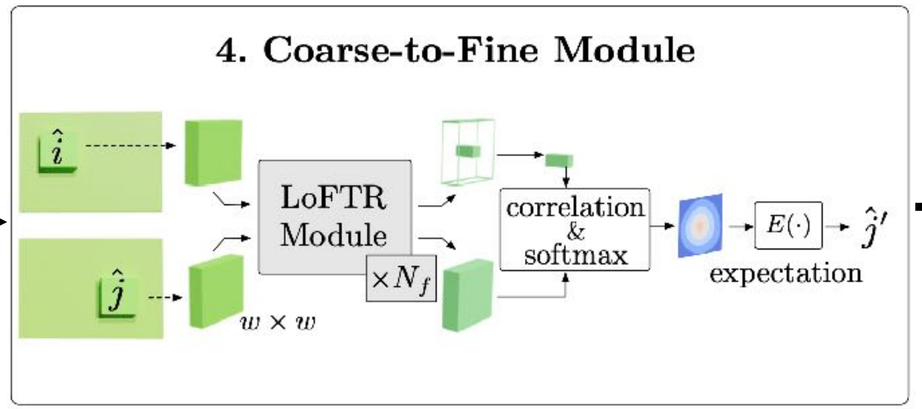
Coarse-to-fine module





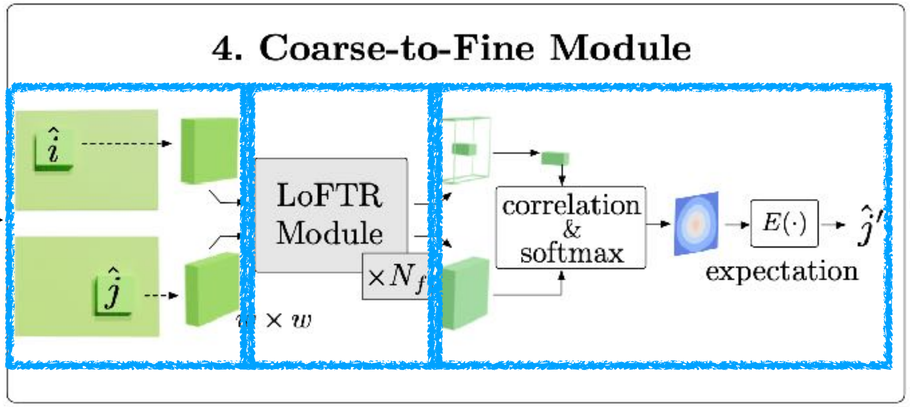
LoFTR

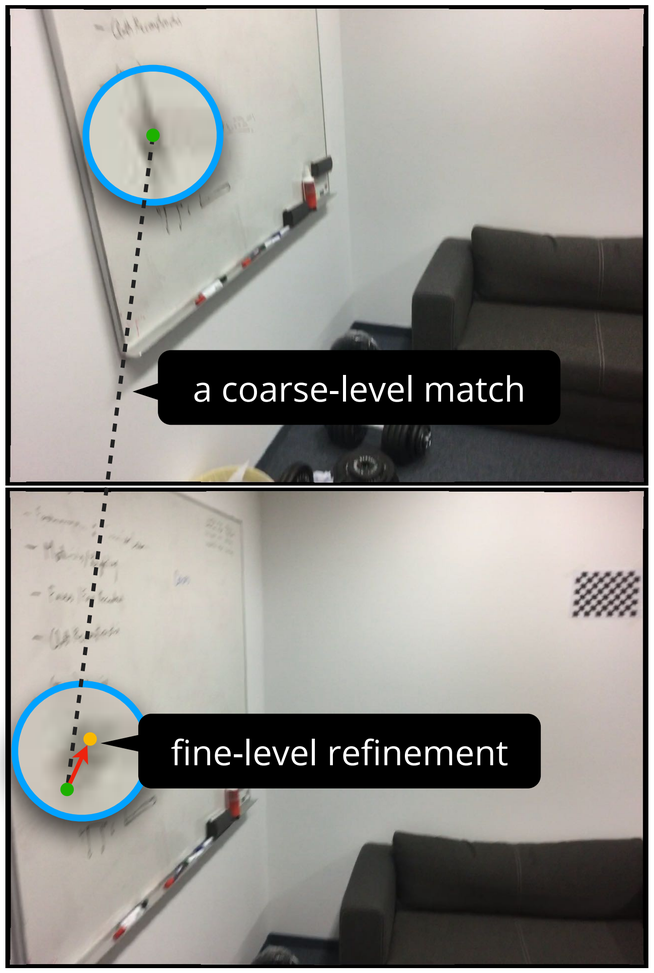
Coarse-to-fine module



LoFTR

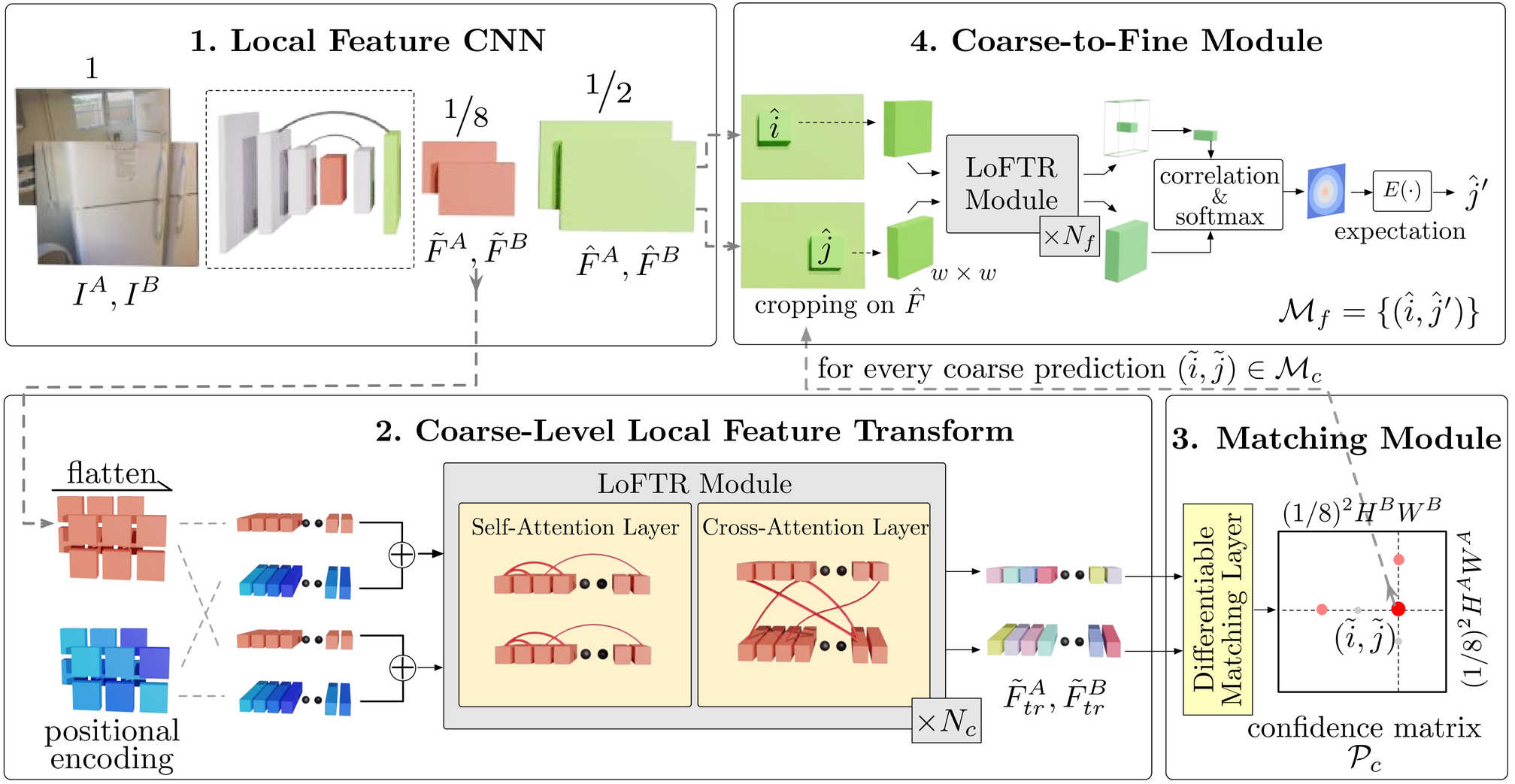
Coarse-to-fine module





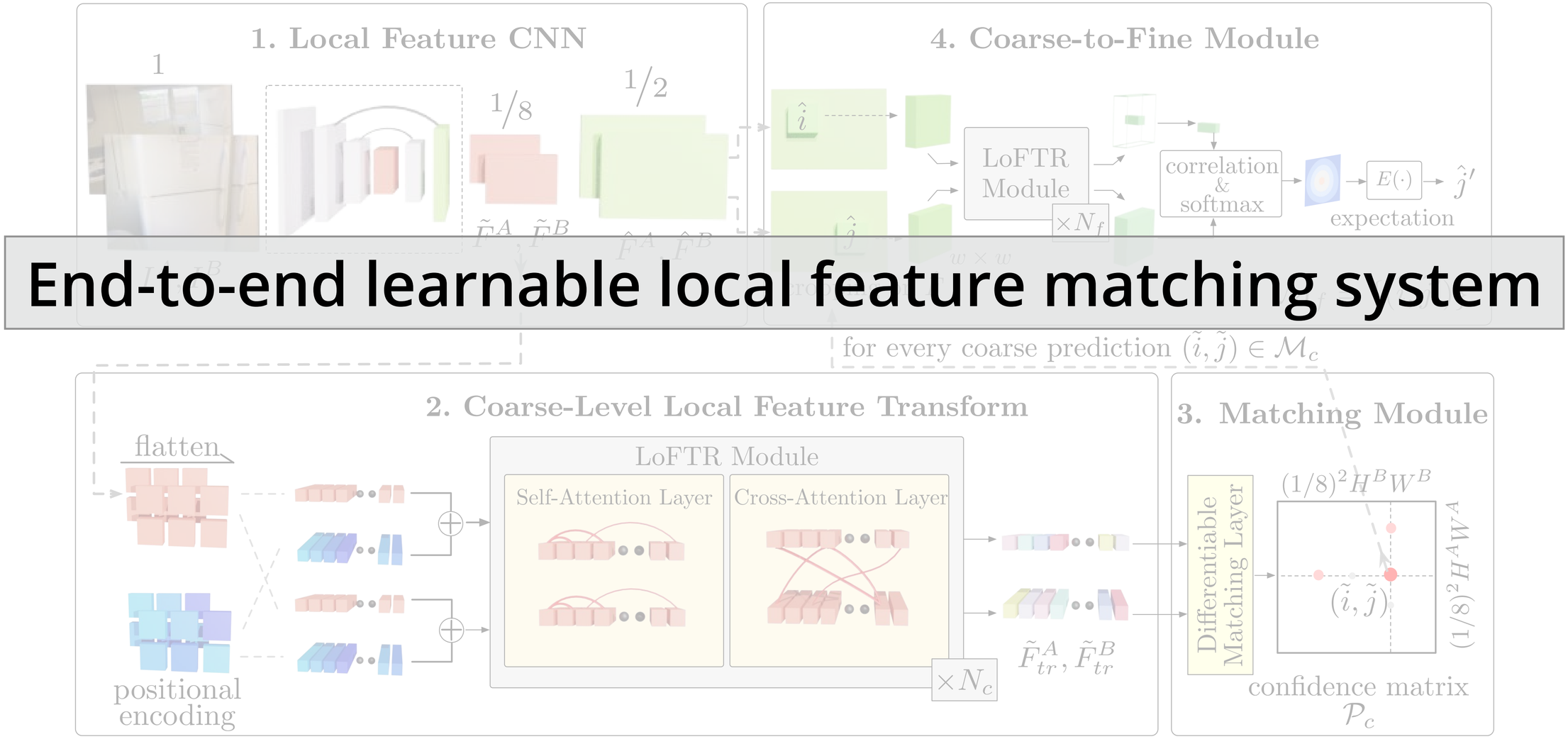
LoFTR

Architecture



LoFTR

Architecture

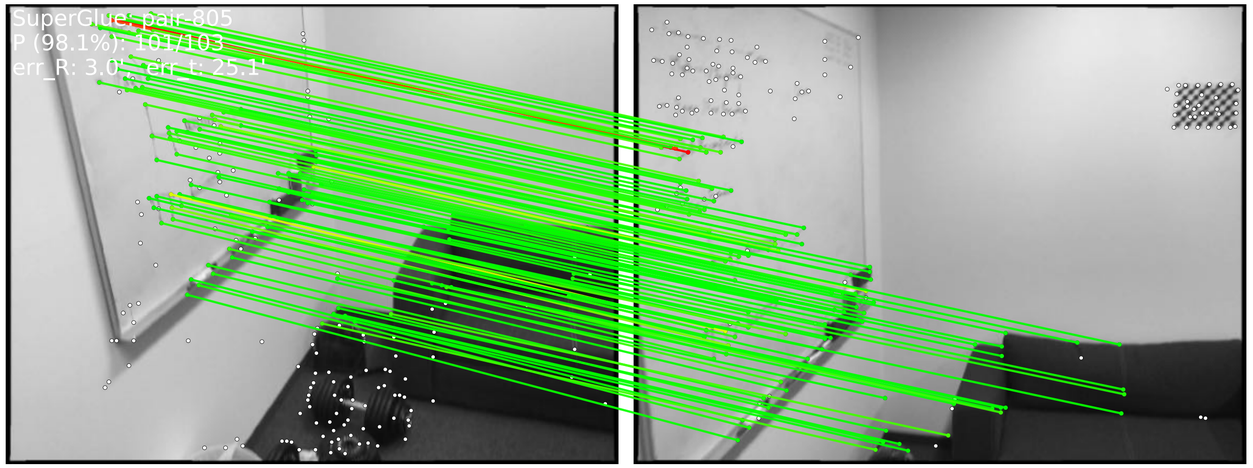


End-to-end learnable local feature matching system

LoFTR

Comparison with SOTA detector-based method

SuperGlue

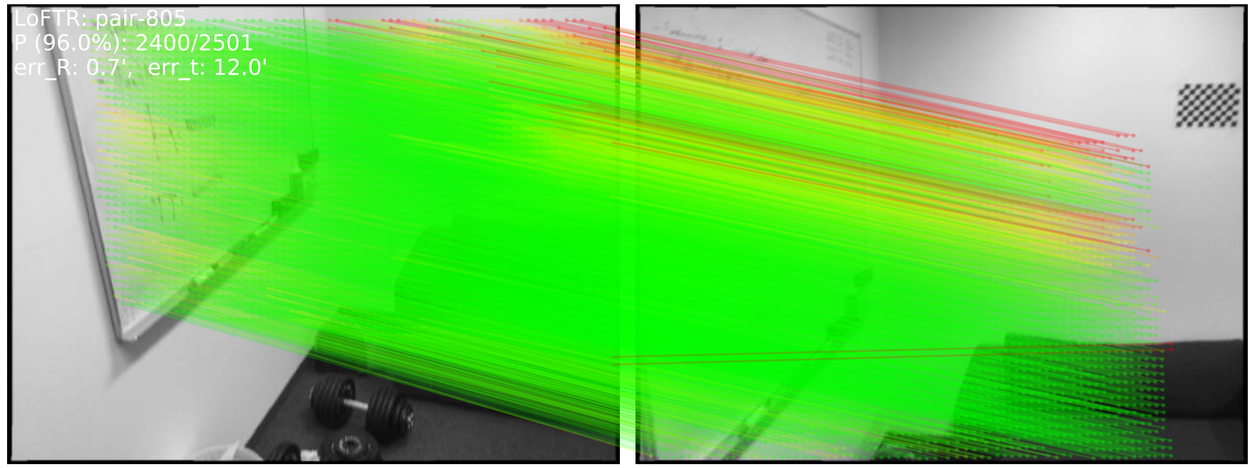


Pose error: $\Delta R = 3.0°, \Delta t = 25.1°$

😃 Sparse matches

😕 Textured regions only

LoFTR



Pose error: $\Delta R = 0.7°, \Delta t = 12.0°$

😃 Semi-dense matches

😃 Textured & Texture-less regions

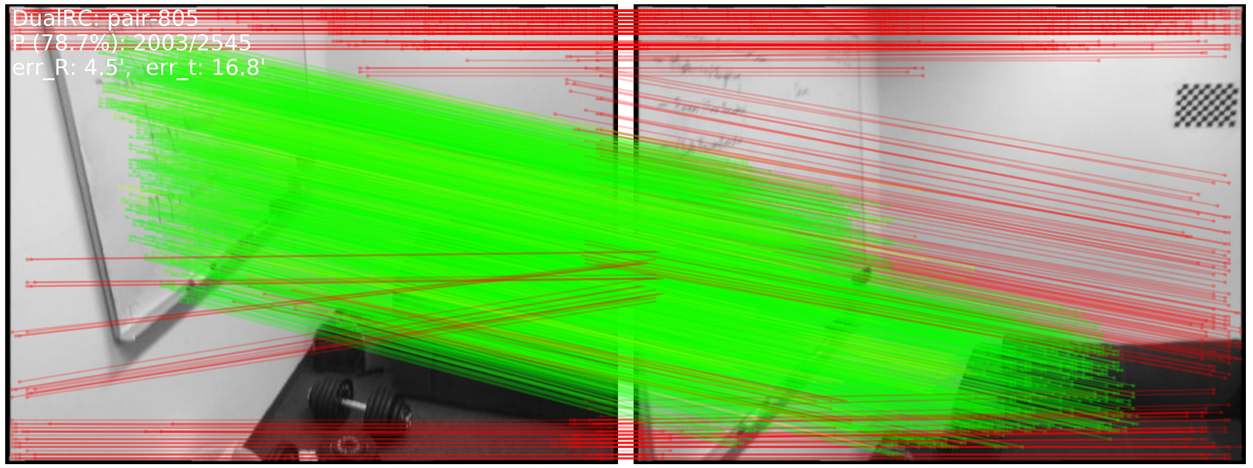
😃 Well-distributed matches

Sarlin, et al. "**SuperGlue: Learning feature matching with graph neural networks**", CVPR 2020

LoFTR

Comparison with SOTA detector-free method

DRC-Net



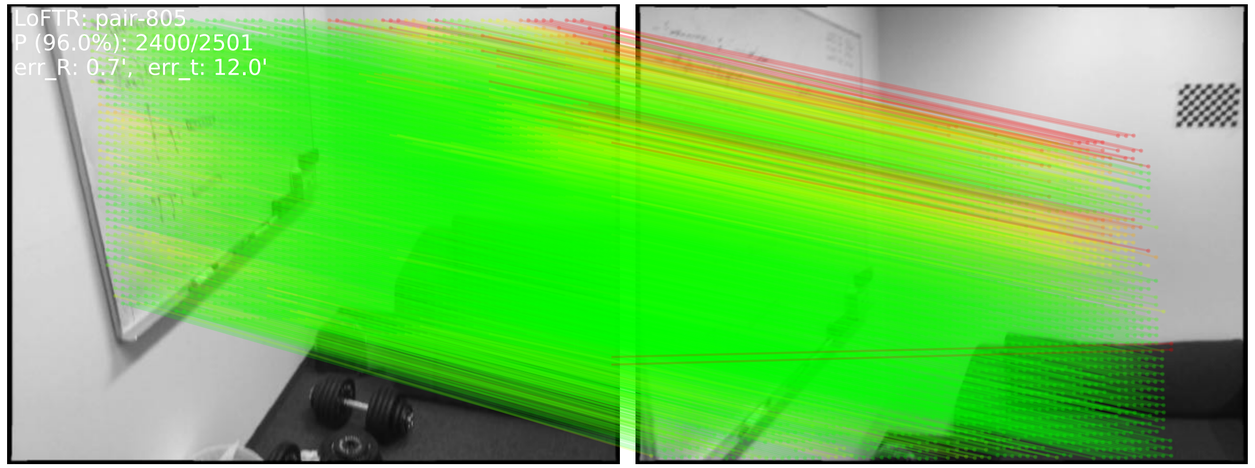
Pose error: $\Delta R = 4.5°, \Delta t = 16.8°$

😕 Local consensus only

👉 many incorrect matches

😕 No matches on purely texture-less regions

LoFTR



Pose error: $\Delta R = 0.7°, \Delta t = 12.0°$

😃 Global consensus

👉 mostly correct matches

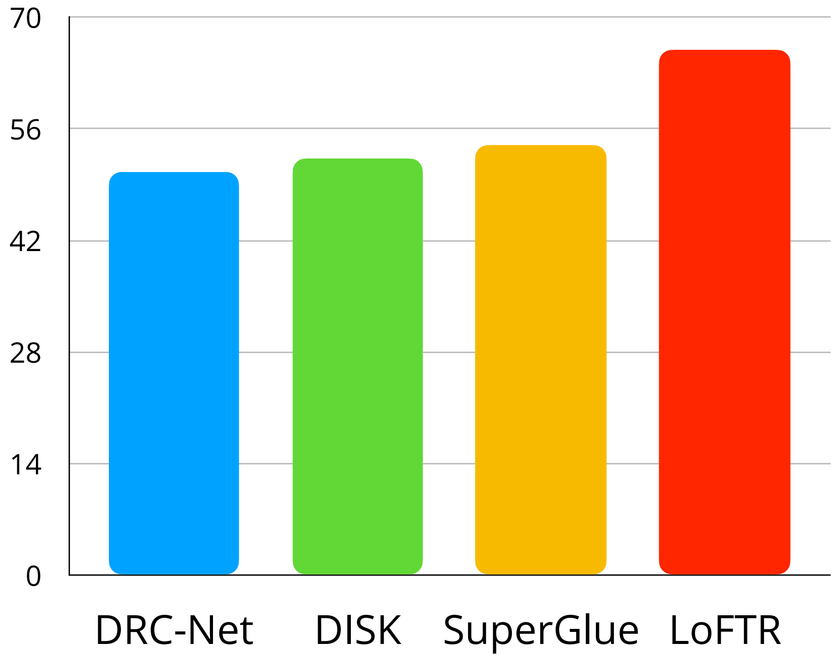
😃 Semi-dense matches on purely texture-less regions

Li, et al. "**Dual-resolution correspondence networks**", NeurIPS 2020

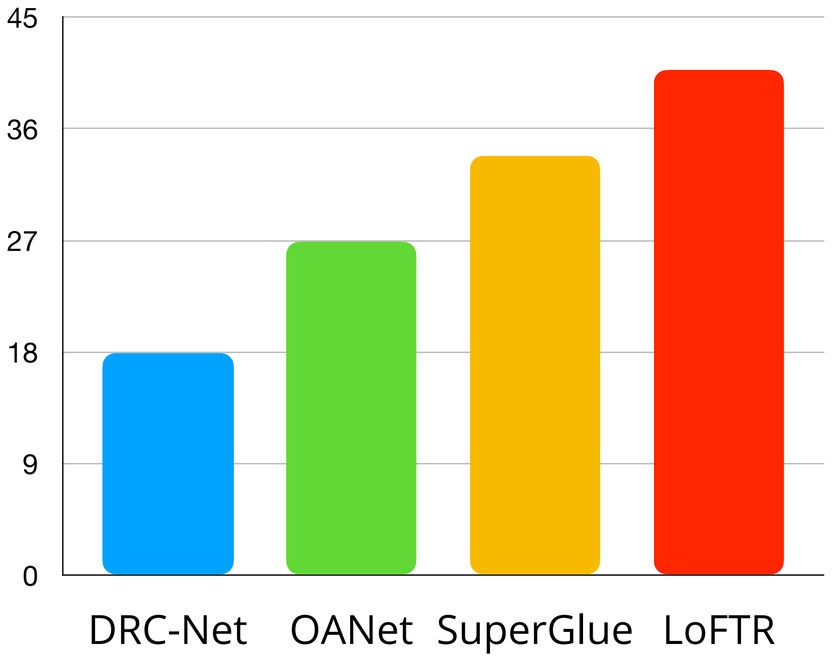
Experiments

Relative pose estimation

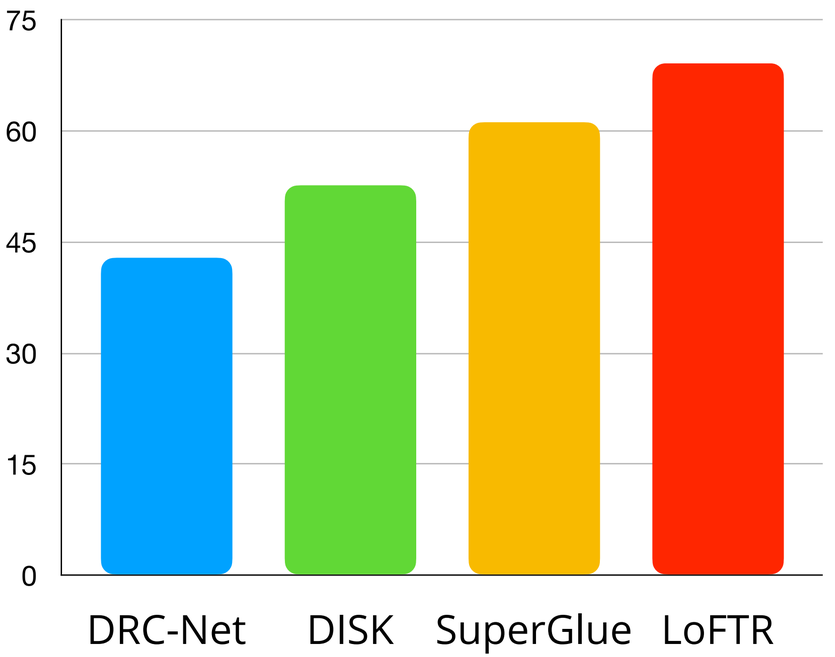
HPatches (AUC@3px)



ScanNet (AUC@10°)

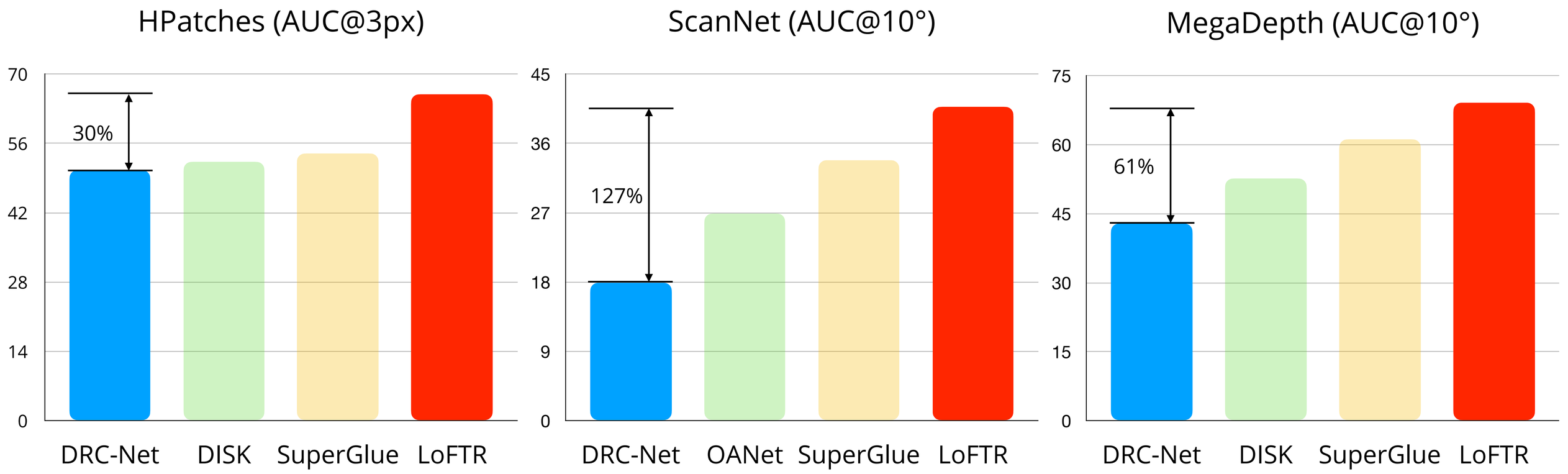


MegaDepth (AUC@10°)



Experiments

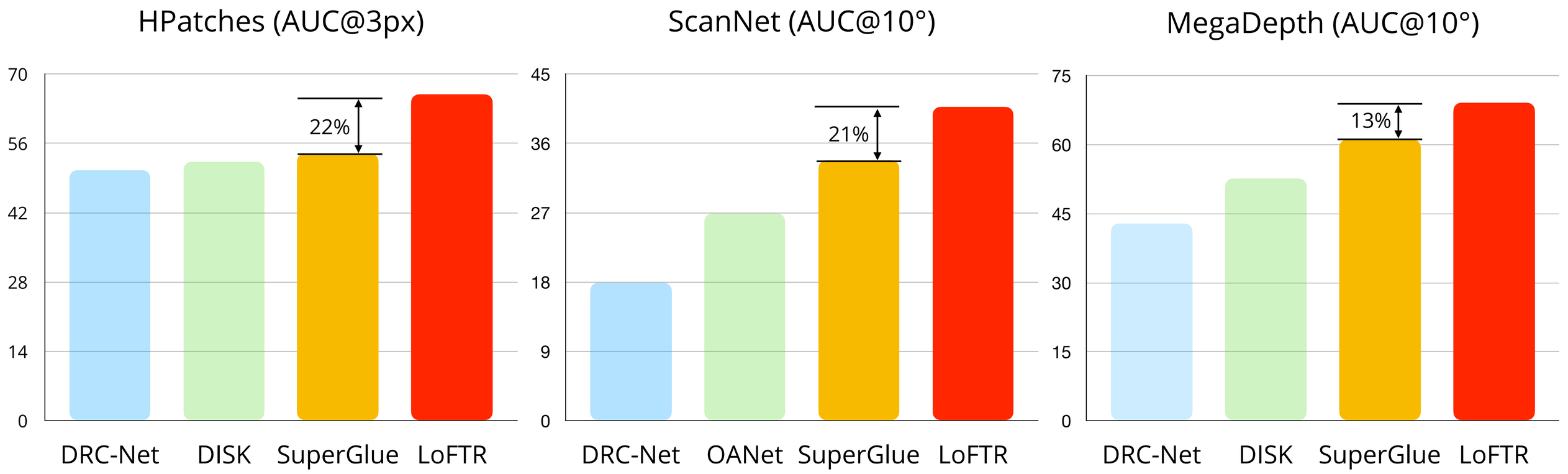
Relative pose estimation



Improvements (in ratio) compared to the SOTA detector-free baseline DRC-Net

Experiments

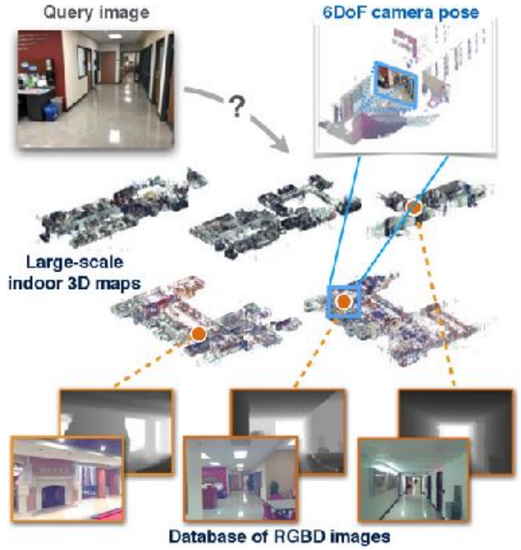
Relative pose estimation



Improvements (in ratio) compared to the SOTA detector-based baseline SuperGlue

Experiments

Visual localization



Rank 🥇 on two VisLoc benchmarks* !

VisLoc Local Feature Challenge (Aachen)

	Night			RANK
	0.25m, 2° / 0.5m, 5° / 1.0m, 10°			
LoFTR	72.8	**88.5**	**99**	🥇
SuperGlue	**73.3**	88.0	98.4	🥈
R2D2	71.2	86.9	98.9	🥉

VisLoc Indoor Localization (InLoc)

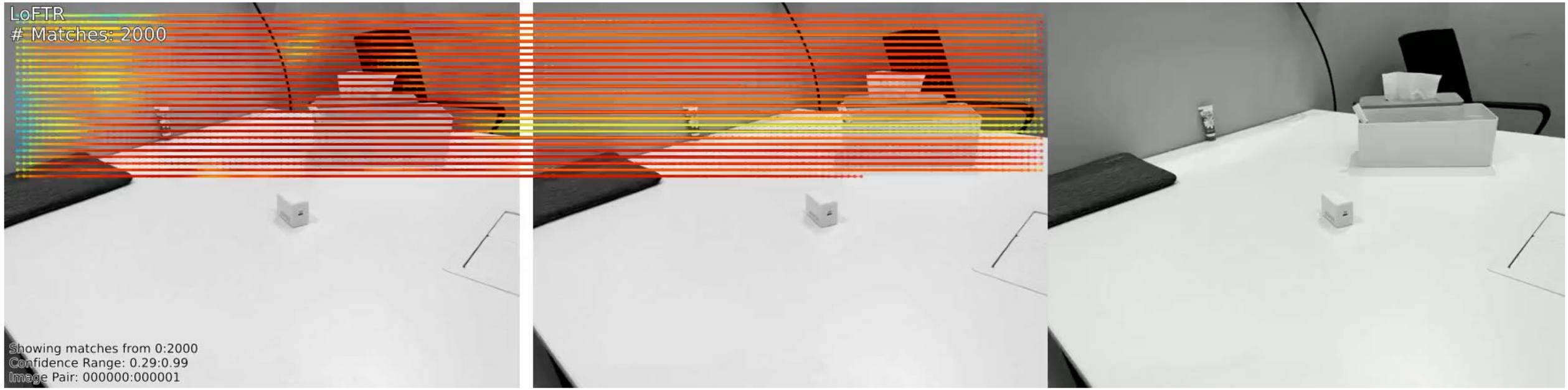
	DUC1			DUC2			RANK
	0.25m, 2° / 5° / 10°			0.25m, 2° / 5° / 10°			
LoFTR + HLOC	47.5	**72.2**	**84.8**	**54.2**	74.8	**85.5**	🥇
SuperGlue + HLOC	**49.0**	68.7	80.8	53.4	**77.1**	82.4	🥈
R2D2 + KAPTURE	41.4	60.1	73.7	47.3	67.2	73.3	🥉

Revaud, et al. "**R2D2: repeatable and reliable detector and descriptor**", NeurIPS 2019

*: among published methods

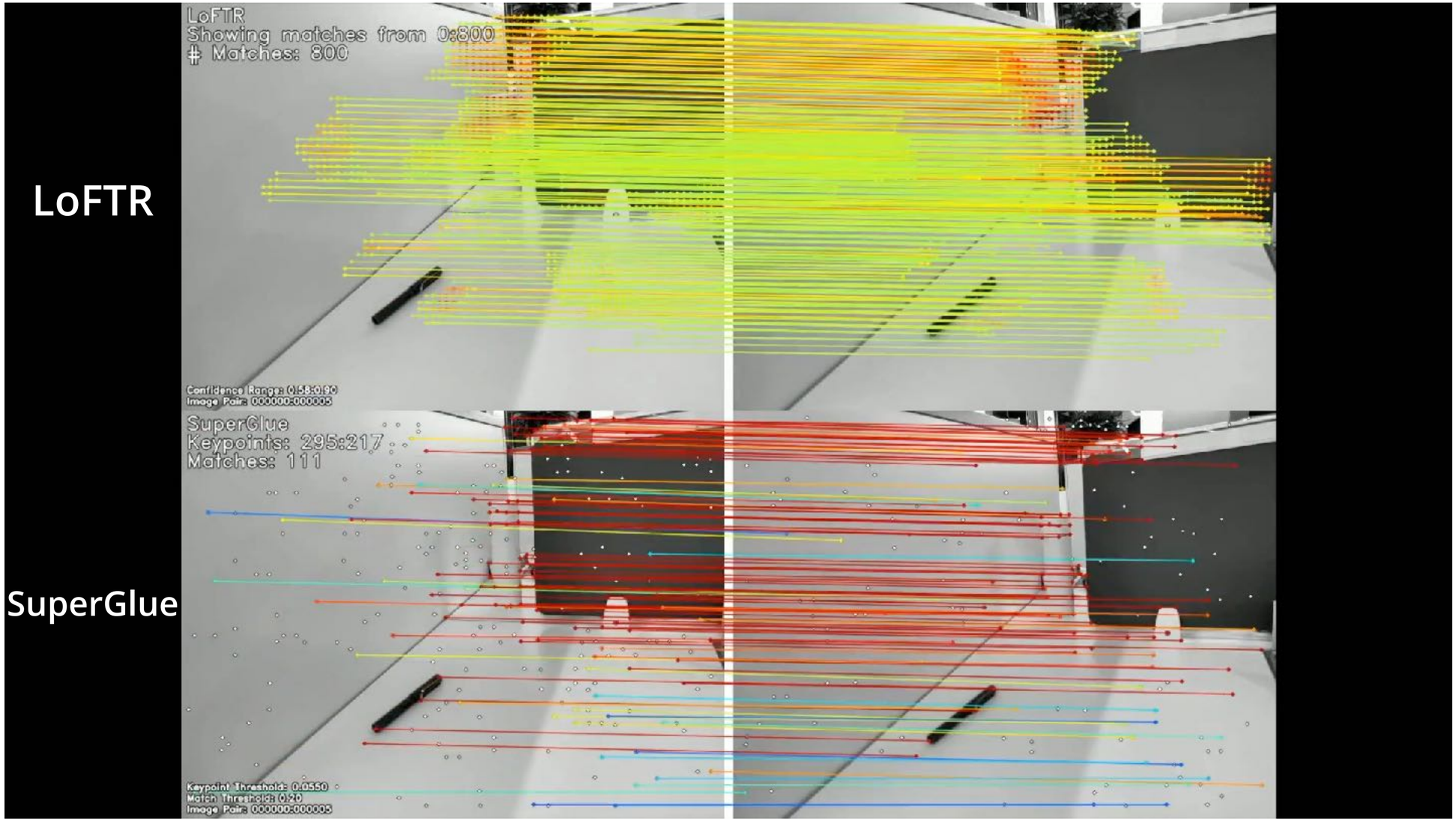
Demo

Matching on a low-texture area



Demo

Comparison with SuperGlue



LoFTR: Detector-Free Local Feature Matching with Transformers

Jiaming Sun* Zehong Shen* Yu'ang Wang* Hujun Bao Xiaowei Zhou

CVPR 2021

Project page: https://zju3dv.github.io/loftr

Code: https://github.com/zju3dv/LoFTR

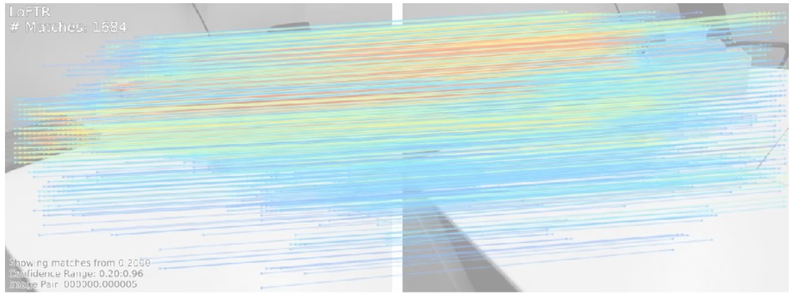
Paper link: https://arxiv.org/pdf/2104.00680.pdf

Summary on Localization

- **Learning to find correspondences instead of directly regress poses**
  - Leverage the geometric structure of the problem (BA, PnP, RANSAC)
  - Let CNNs/Transformers do what they do best (learning feature representations)
- **Learning feature matching**
  - Looking at both image at same time, instead of detecting interests points independently, learns position-dependent feature representations
  - Leverage the inductive bias of different network architectures (CNNs: local, Transformers: global)
  - Learn the matches end-to-end directly

Today's Topic

Visual localization

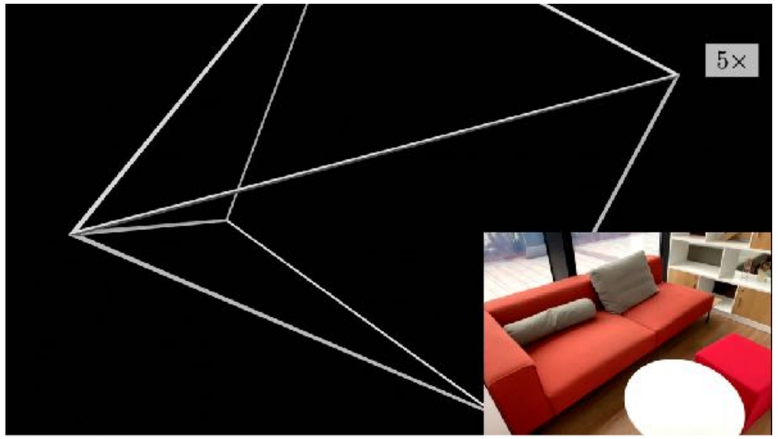


LoFTR: Detector-Free Local Feature Matching with Transformers

Jiaming Sun* Zehong Shen* Yu'ang Wang* Hujun Bao Xiaowei Zhou

CVPR 2021

3D Reconstruction



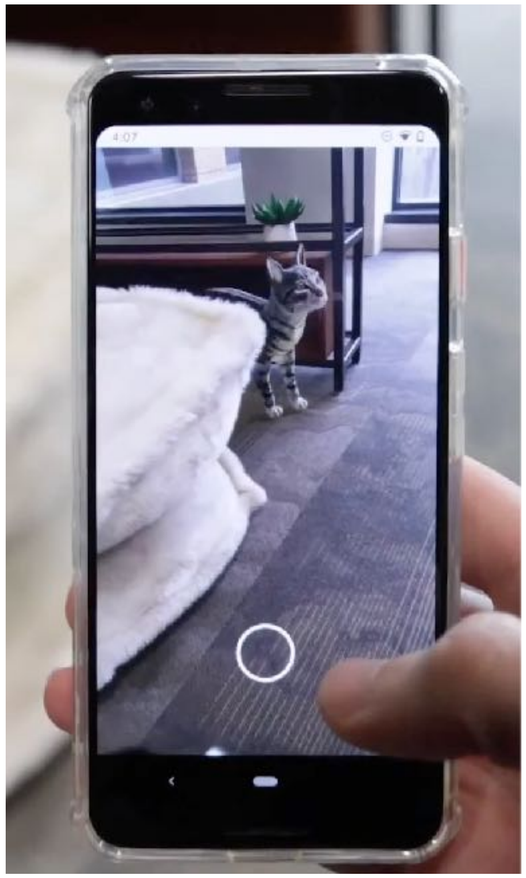
NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video

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

CVPR 2021 (Oral)

Motivation

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



Credit: DepthLab, Apple Clips with LiDAR, 3D Scanner App

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

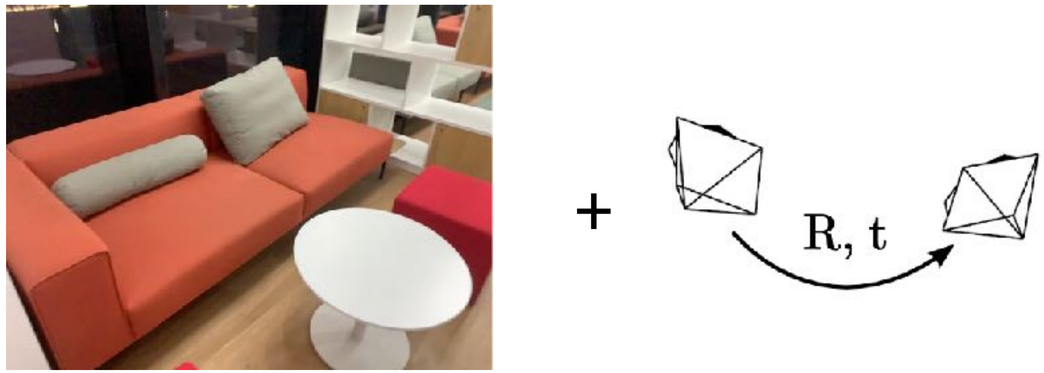
With a depth sensor

😃 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

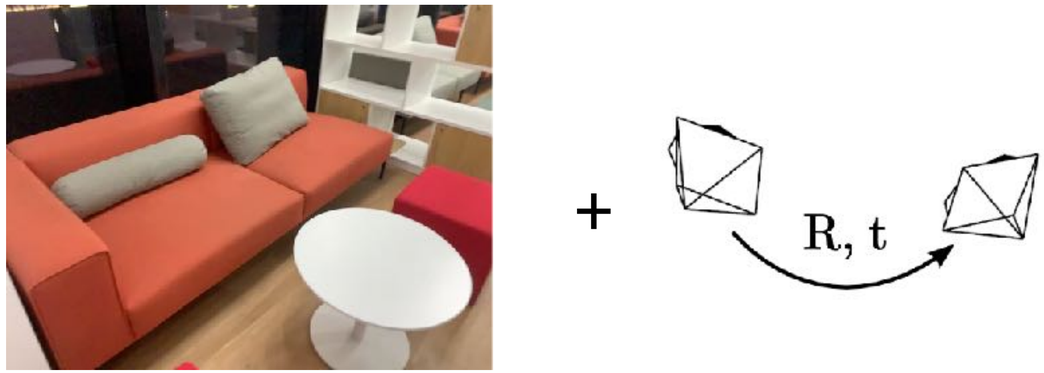
With a depth sensor

😀 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

😕 Poor 3D reconstruction quality

Credit: RealSense, Azure Kinect, LiDAR on iPad Pro

Motivation

Depth sensor v.s. Monocular camera

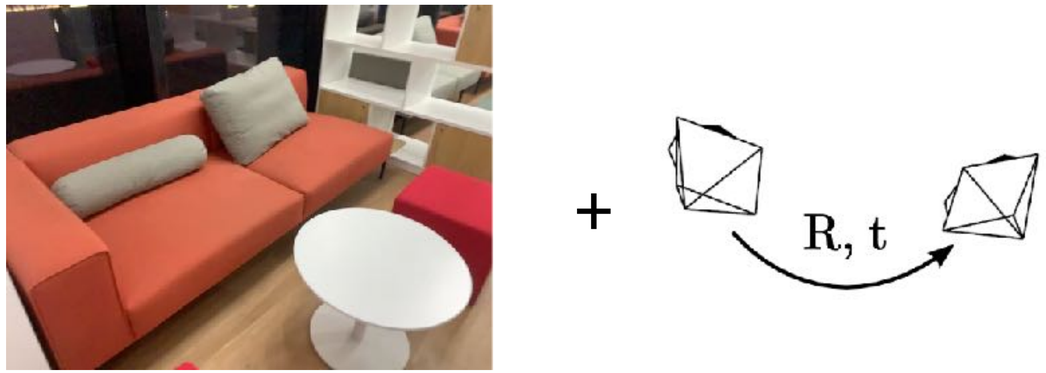
With a depth sensor

😀 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

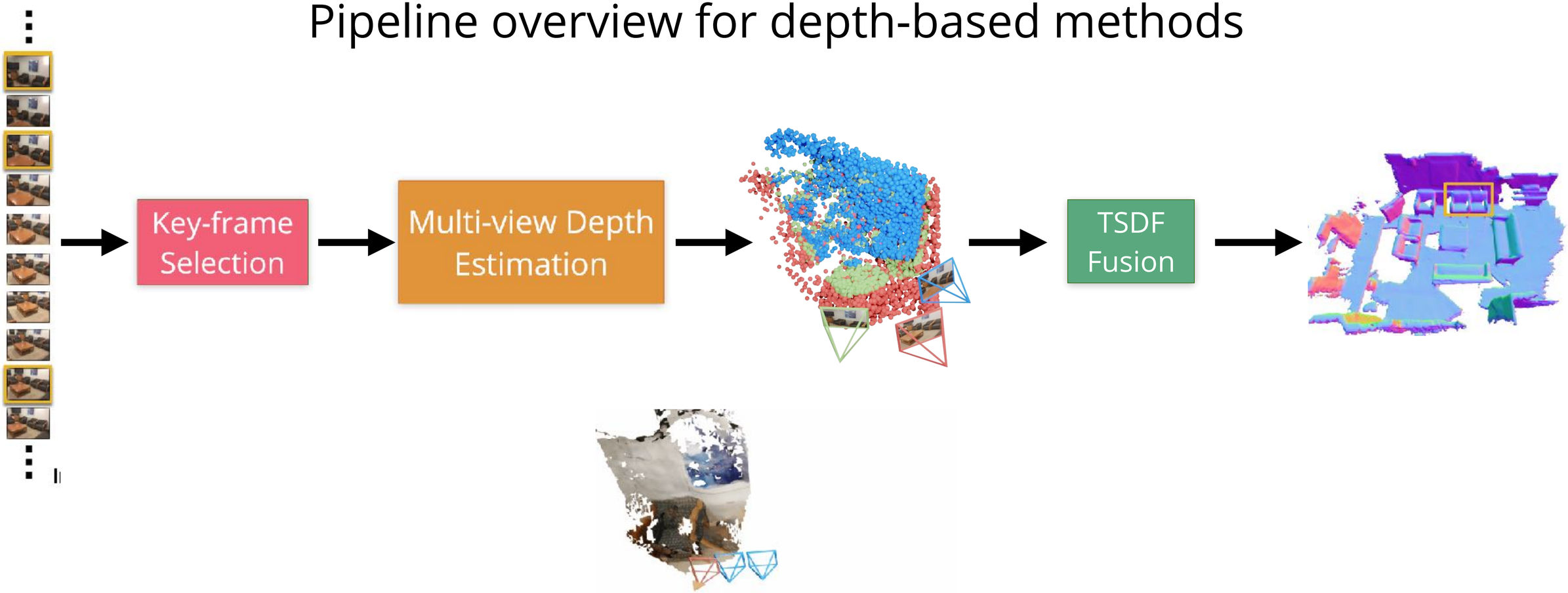
😕 Not as accurate as depth sensors

😕 Poor 3D reconstruction quality

Credit: RealSense, Azure Kinect, LiDAR on iPad Pro

Motivation

Pipeline overview for depth-based methods

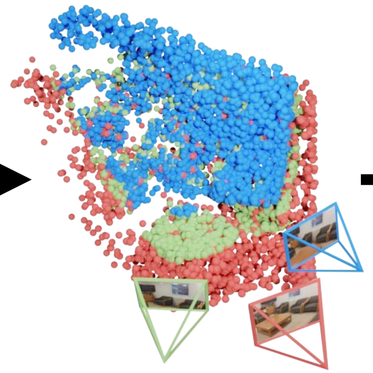


Motivation

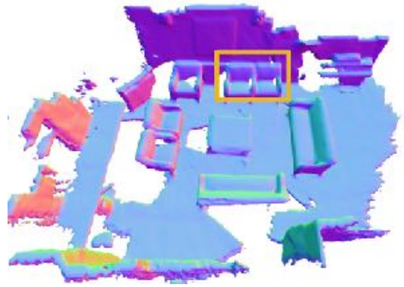
Pipeline overview for depth-based methods

Key-frame Selection

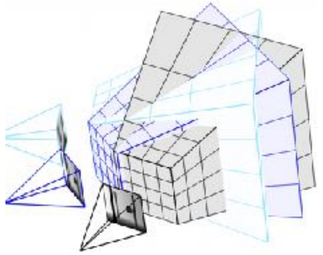
Multi-view Depth Estimation



TSDF Fusion



Depth Sensor



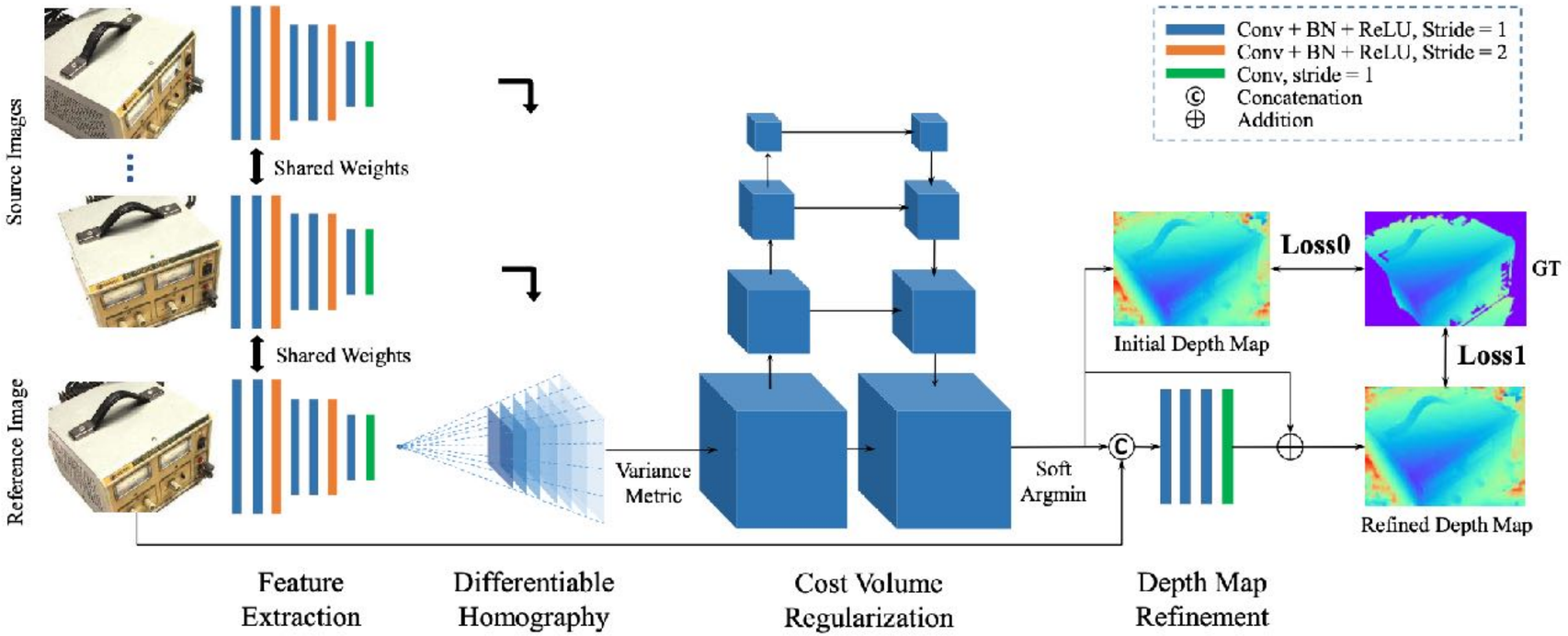
KinectFusion

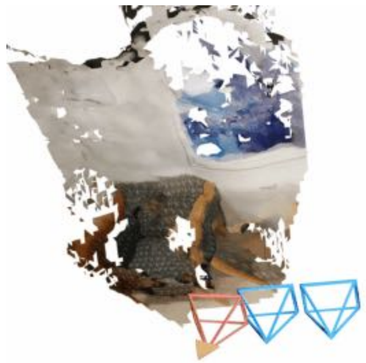
Newcombe et al., "**KinectFusion: Real-Time Dense Surface Mapping and Tracking**", *UIST 2011*

Motivation

MVSNet: Depth Inference for Unstructured Multi-view Stereo

Yao 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

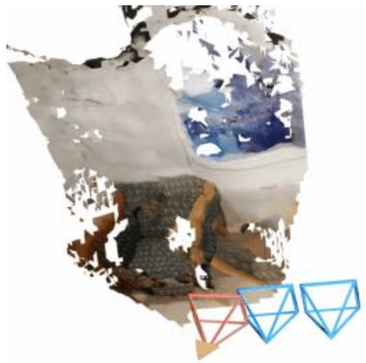
Zachary Teed
Princeton University
zteed@cs.princeton.edu

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

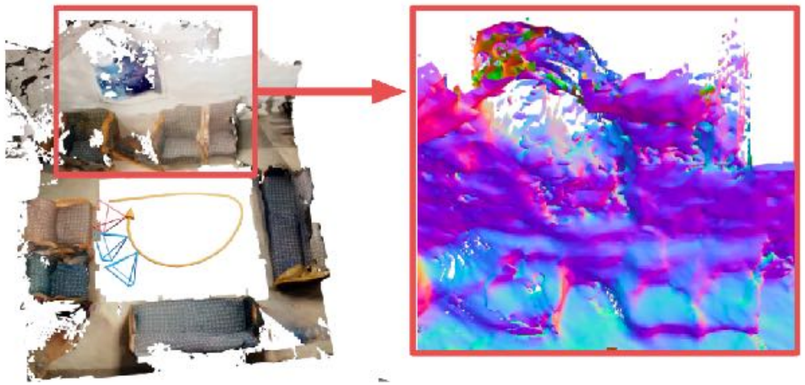
Chao Liu[1,2*] Jinwei Gu[1,3*] Kihwan Kim[1] Srinivasa Narasimhan[2] Jan Kautz[1]
[1]NVIDIA [2]Carnegie Mellon University [3]SenseTime

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

Motivation

Depth-based methods v.s. NeuralRecon

Depth-based methods

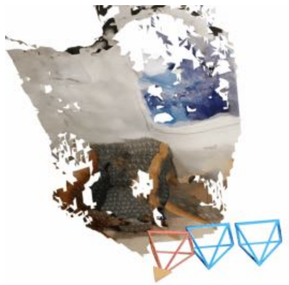


😕 Either layered or scattered results

Motivation

Depth-based methods v.s. NeuralRecon

Depth-based methods



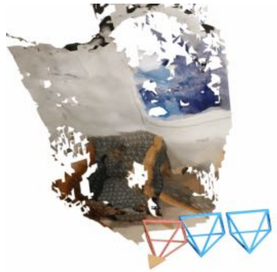
🙁 Either layered or scattered results

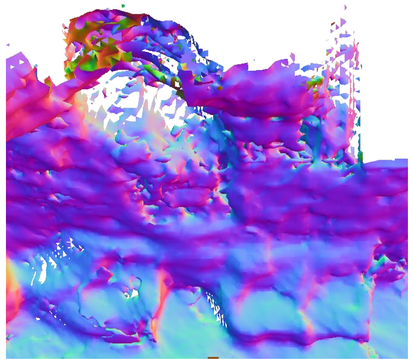
😕 Redundant computation

Motivation

Depth-based methods v.s. NeuralRecon

Depth-based methods

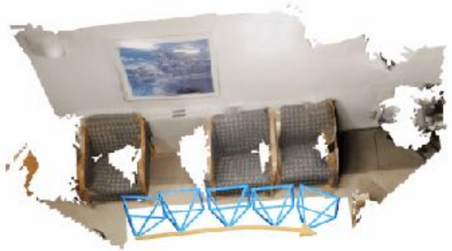


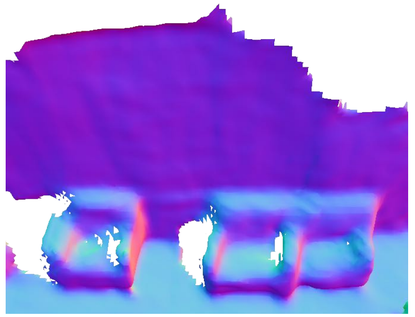


😕 Either layered or scattered results

😕 Redundant computation

Our solution: NeuralRecon





😃 Reconstruct local surfaces directly in TSDF

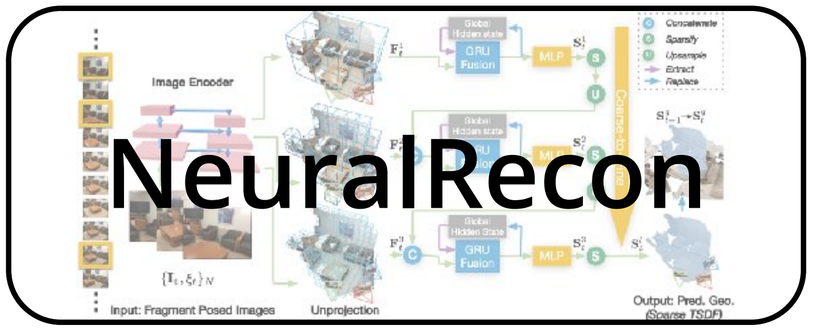
😃 Joint fragment reconstruction and fusion

😃 Better quality and faster speed

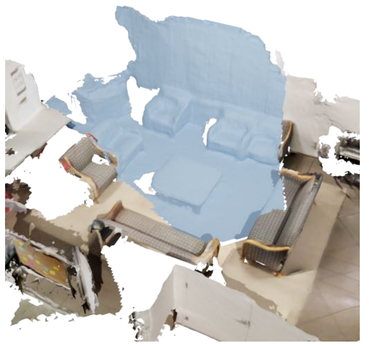
NeuralRecon

Overview

Input:
Posed Images



End-to-End System



Output:
Scene Geometry
(sparse TSDF volume)

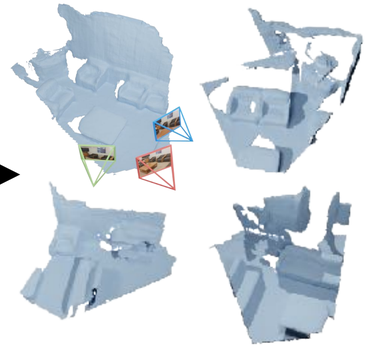
NeuralRecon

Overview

Input:
Posed Images

Key-frame Selection

Fragment Reconstruction



...

Fragment Geometry
(sparse TSDF volume)

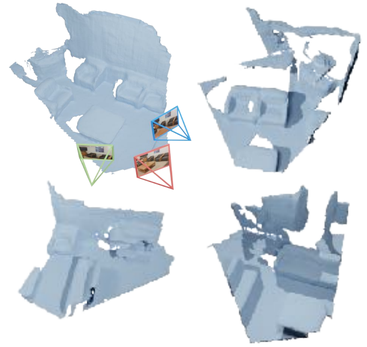
Fuse to Global Volume

Output:
Scene Geometry
(sparse TSDF volume)

NeuralRecon

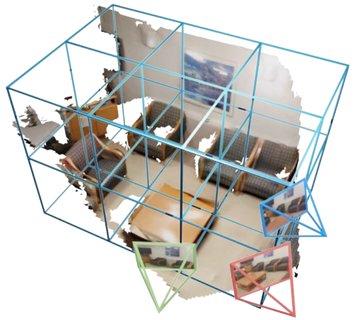
Fragment reconstruction overview

Fragment Reconstruction

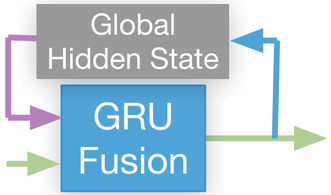


...

Fragment Geometry
(sparse TSDF volume)



View-Independent Volume

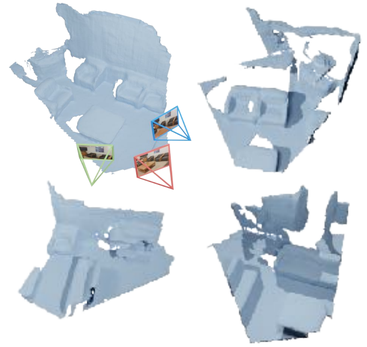


Joint TSDF Reconstruction and Fusion

NeuralRecon

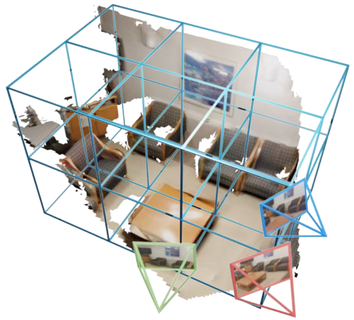
Fragment reconstruction overview

Fragment Reconstruction

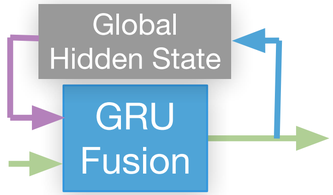


...

Fragment Geometry
(sparse TSDF volume)

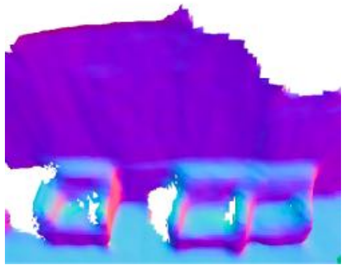


View-Independent Volume



Joint TSDF Reconstruction and Fusion

Real-time

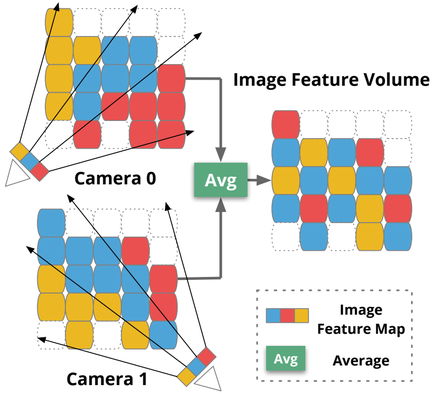


Coherent

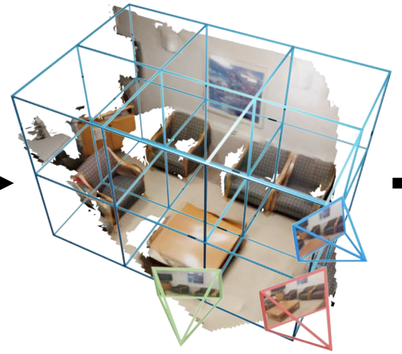
NeuralRecon

Fragment reconstruction

Input:
Posed Images

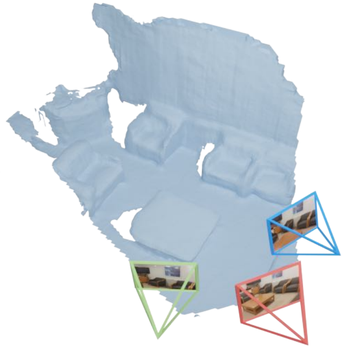


Feature
Unprojection



View-Independent
Volume

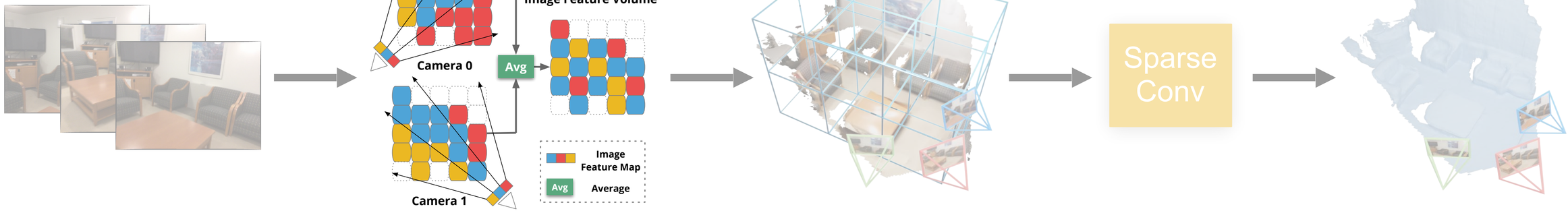
Sparse
Conv



Fragment Geometry
(sparse TSDF volume)

NeuralRecon

Fragment reconstruction



Input:
Posed Images

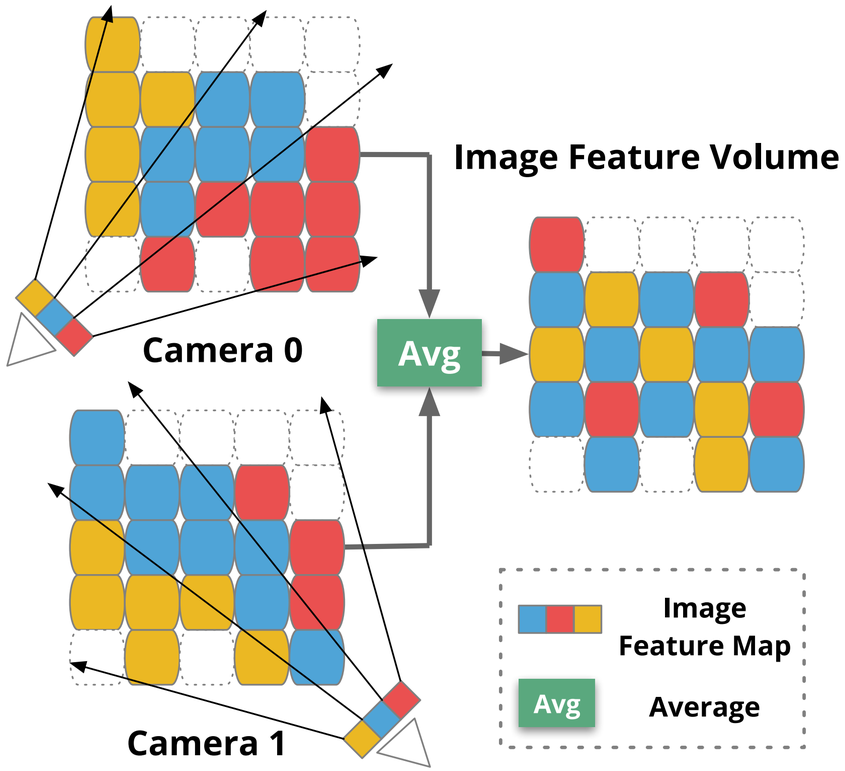
Feature
Unprojection

View-Independent
Volume

Fragment Geometry
(sparse TSDF volume)

NeuralRecon

Fragment reconstruction

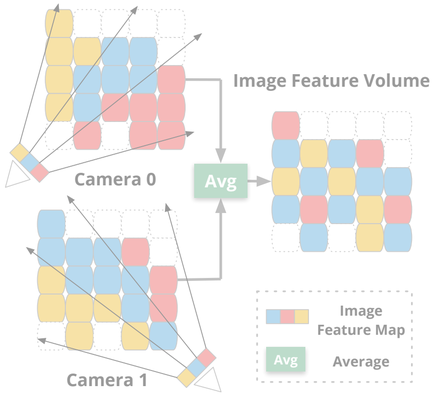


Feature
Unprojection

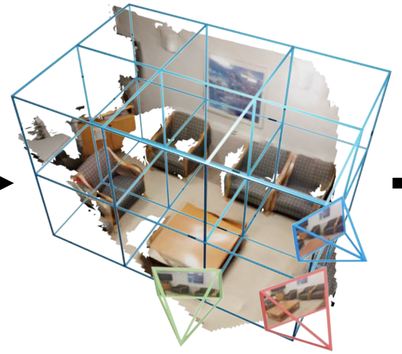
NeuralRecon

Fragment reconstruction

Input:
Posed Images



Feature
Unprojection



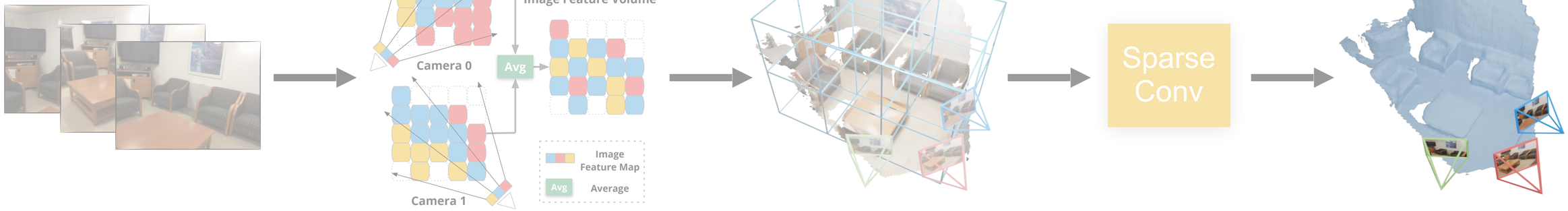
View-Independent
Volume

Sparse
Conv

Fragment Geometry
(sparse TSDF volume)

NeuralRecon

Fragment reconstruction



Input:
Posed Images

Feature
Unprojection

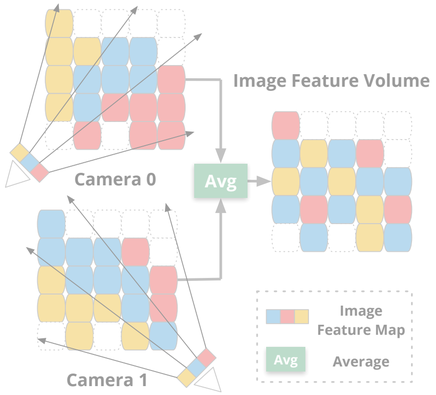
View-Independent
Volume

Fragment Geometry
(sparse TSDF volume)

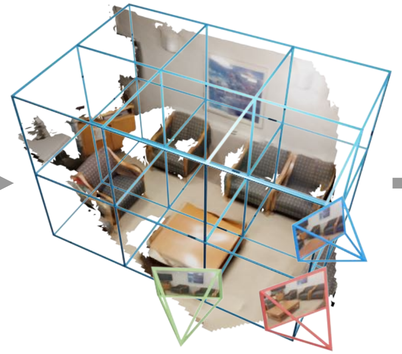
NeuralRecon

Fragment reconstruction

Input:
Posed Images



Feature
Unprojection



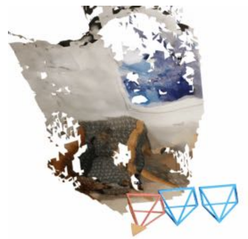
View-Independent
Volume

Sparse
Conv

Fragment Geometry
(sparse TSDF volume)

Why is it better?

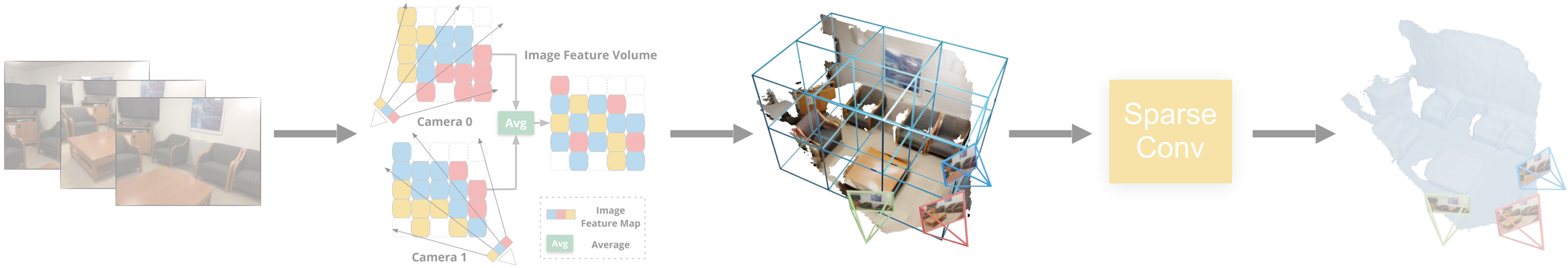
Volume-based



Depth-based

NeuralRecon

Fragment reconstruction



Input: Posed Images

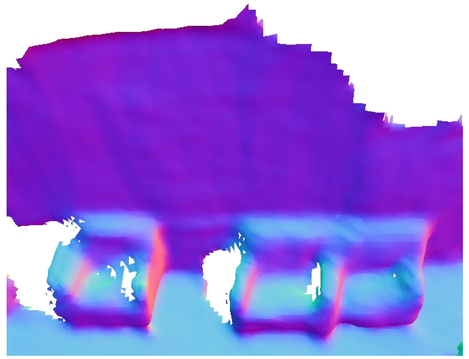
Feature Unprojection

View-Independent Volume

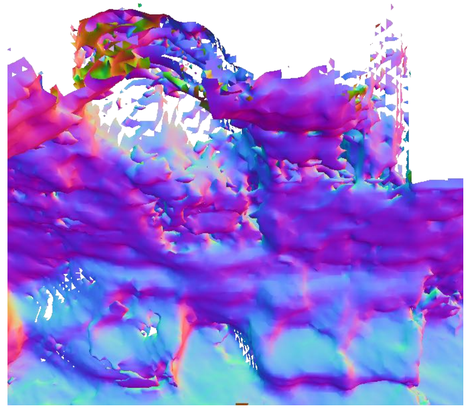
Fragment Geometry *(sparse TSDF volume)*

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

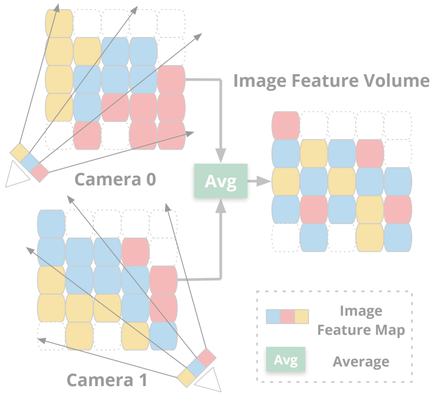


Depth-based

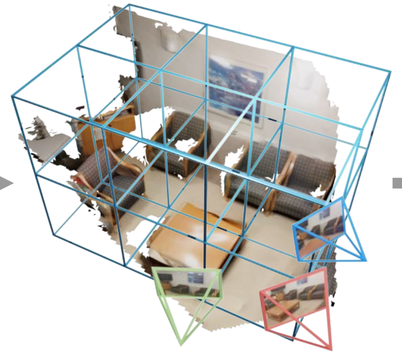
NeuralRecon

Fragment reconstruction

Input:
Posed Images



Feature
Unprojection



View-Independent
Volume

Sparse
Conv

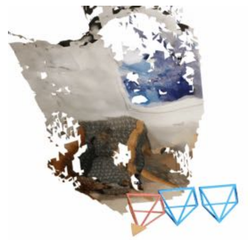
Fragment Geometry
(sparse TSDF volume)

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*

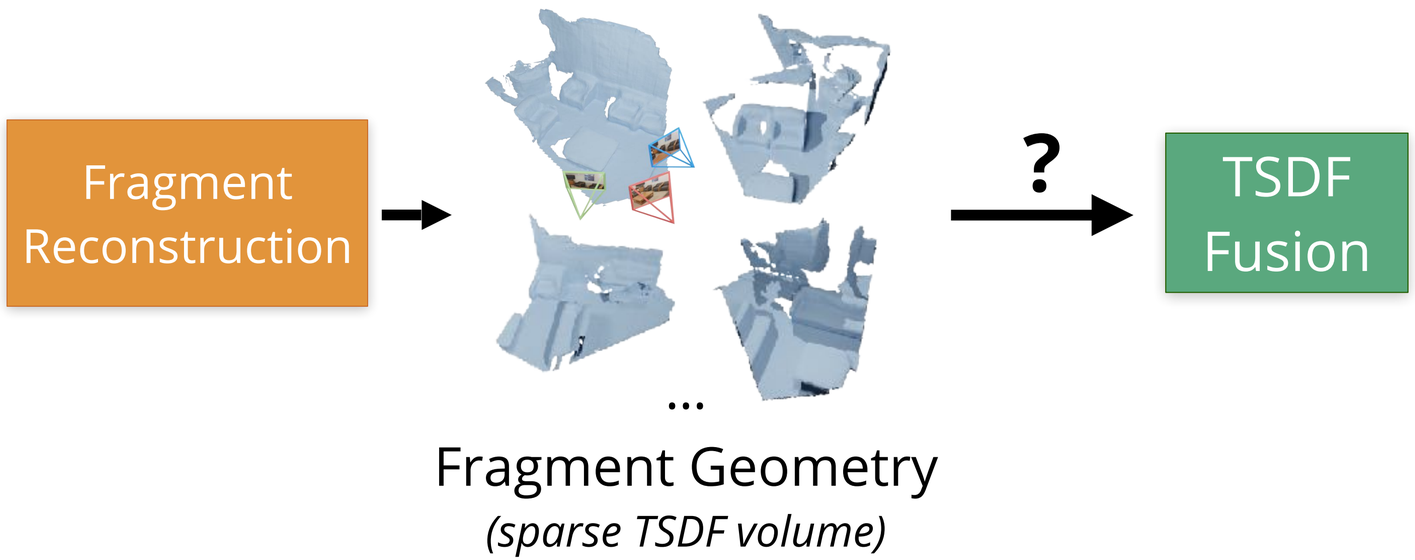
Volume-based



Depth-based

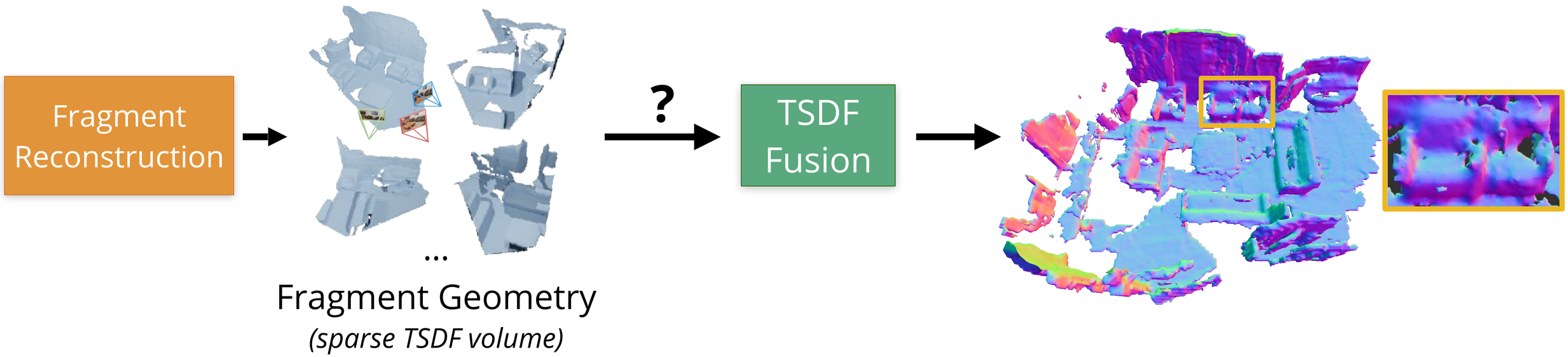
NeuralRecon

TSDF Fusion?



NeuralRecon

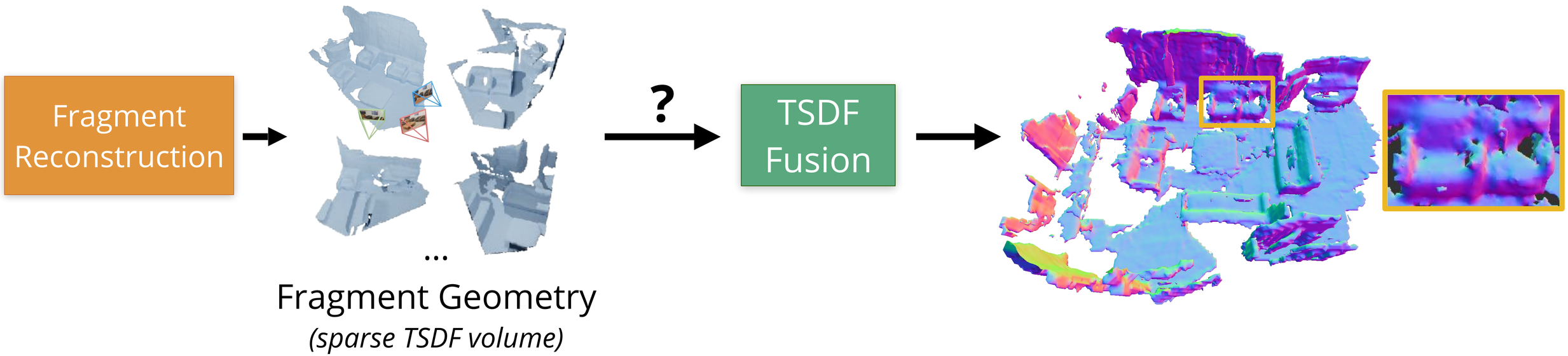
TSDF Fusion?



🙁 **Too many artifacts!**

NeuralRecon

TSDF Fusion?



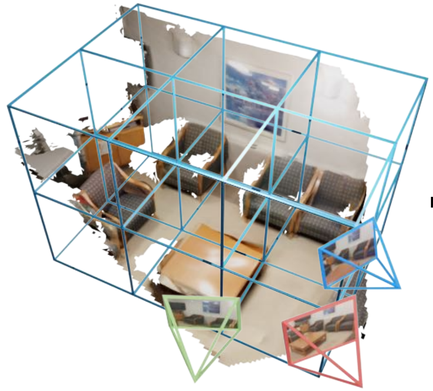
😕 **Too many artifacts!**

Reason: predicted TSDFs are not consistent between fragments

NeuralRecon

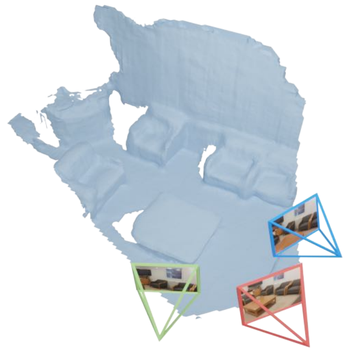
Joint reconstruction and fusion

Input: Posed Images



View-Independent Volume

Sparse Conv

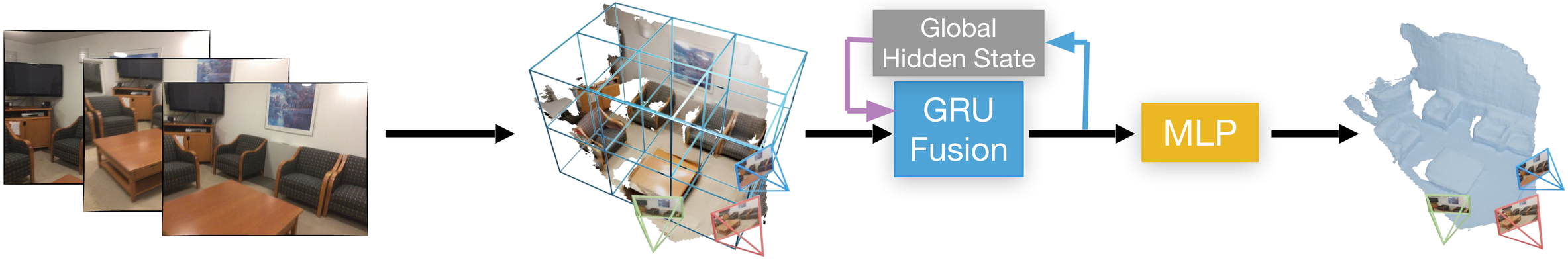


Fragment Geometry

(sparse TSDF volume)

NeuralRecon

Joint reconstruction and fusion



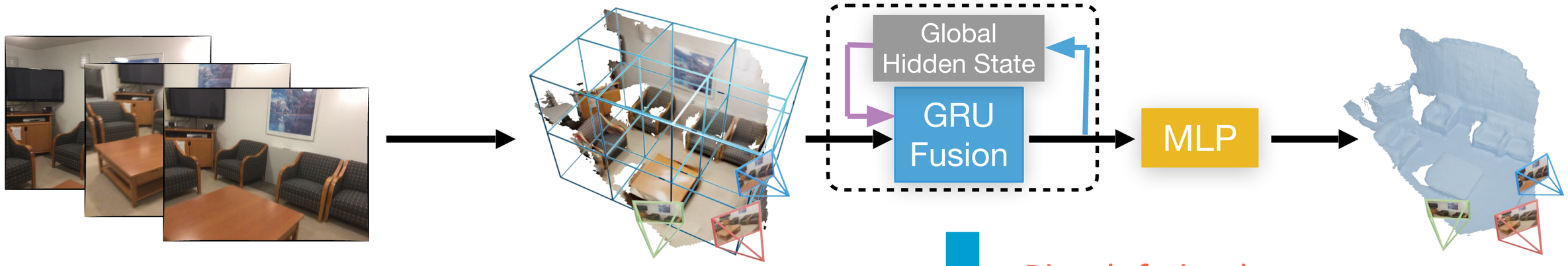
Input: Posed Images

View-Independent
Volume

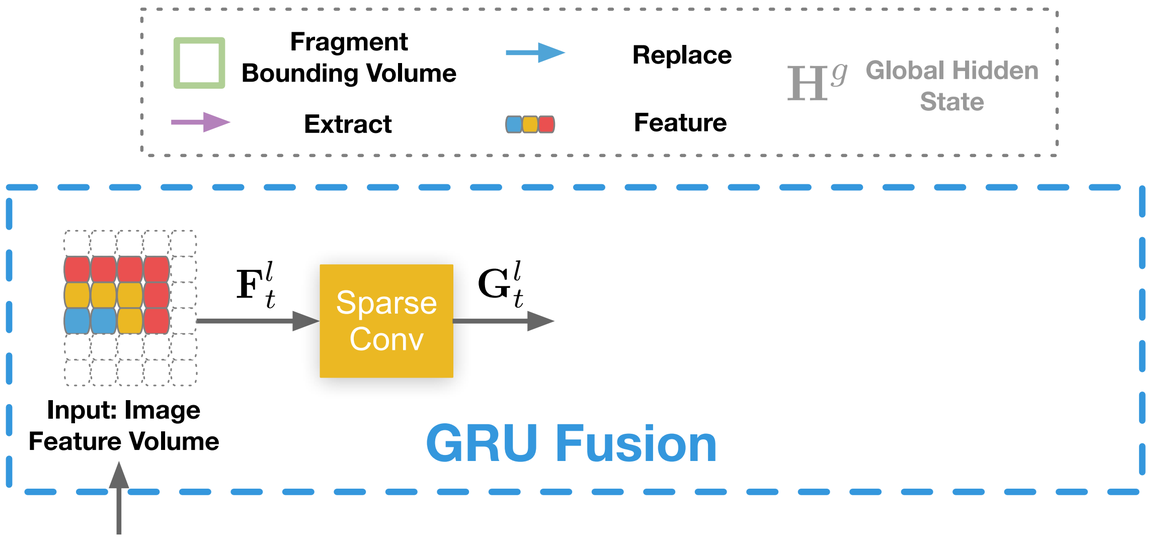
Fragment Geometry
(sparse TSDF volume)

NeuralRecon

Joint reconstruction and fusion

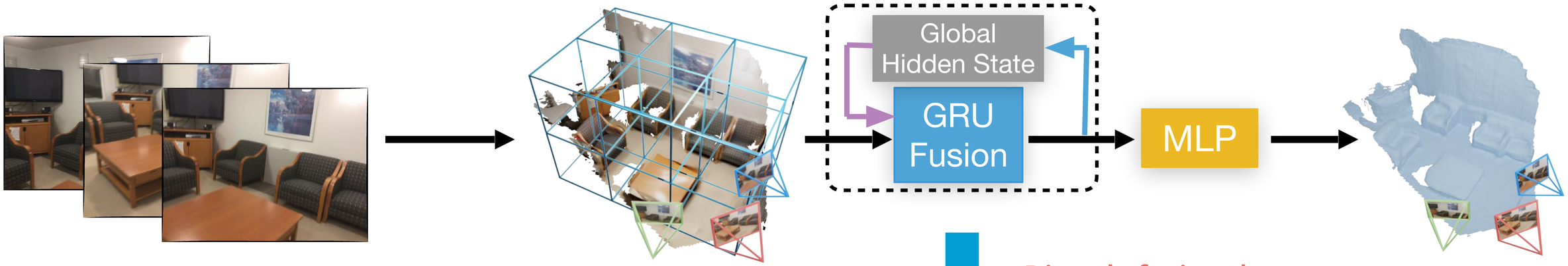


Directly fusing the features with GRU

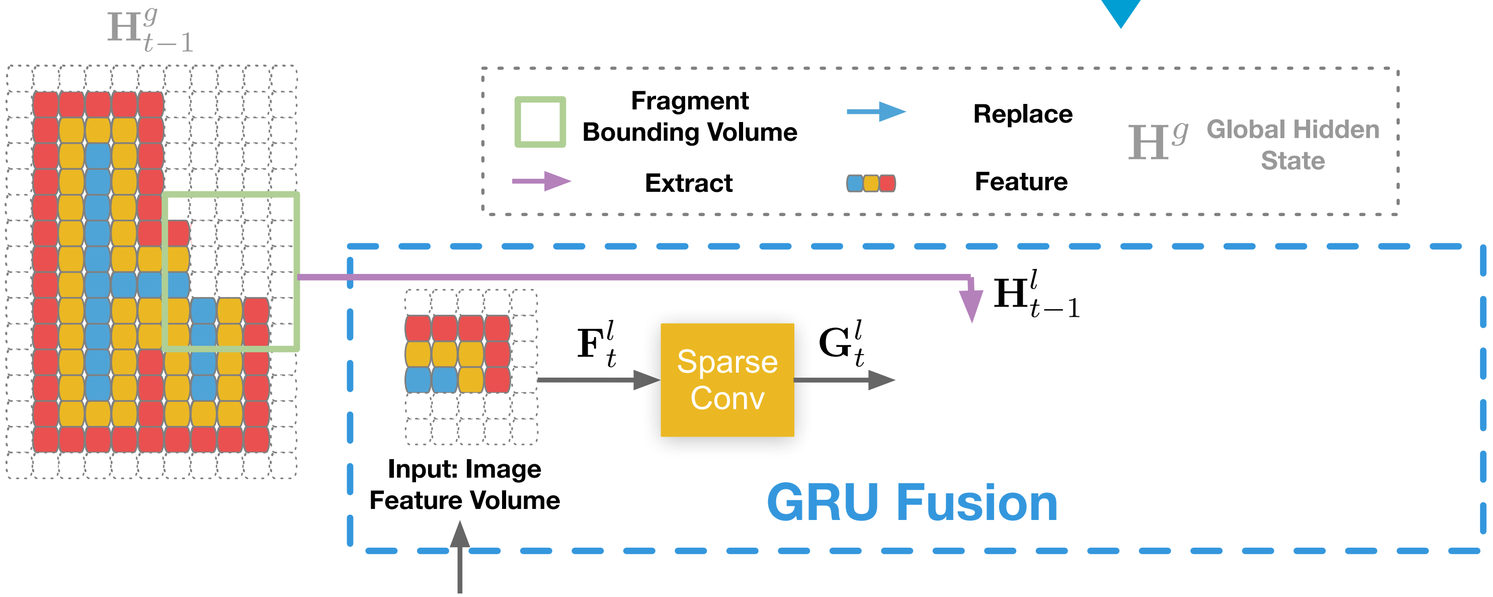


NeuralRecon

Joint reconstruction and fusion

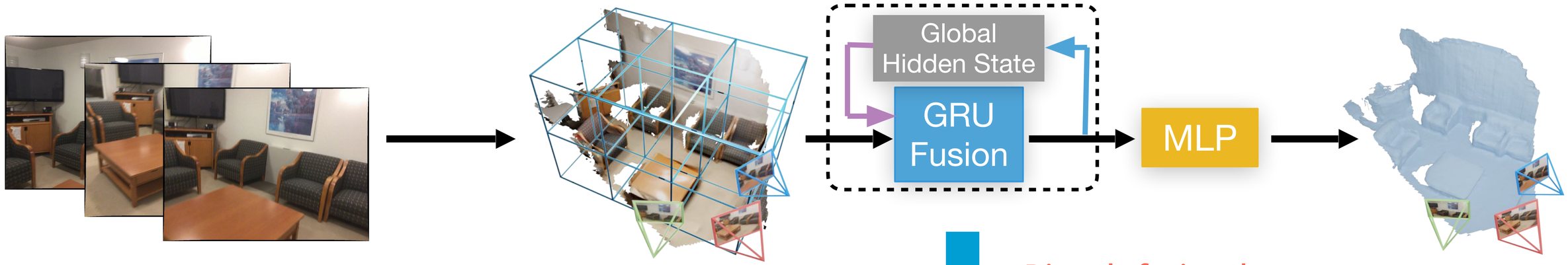


Directly fusing the features with GRU

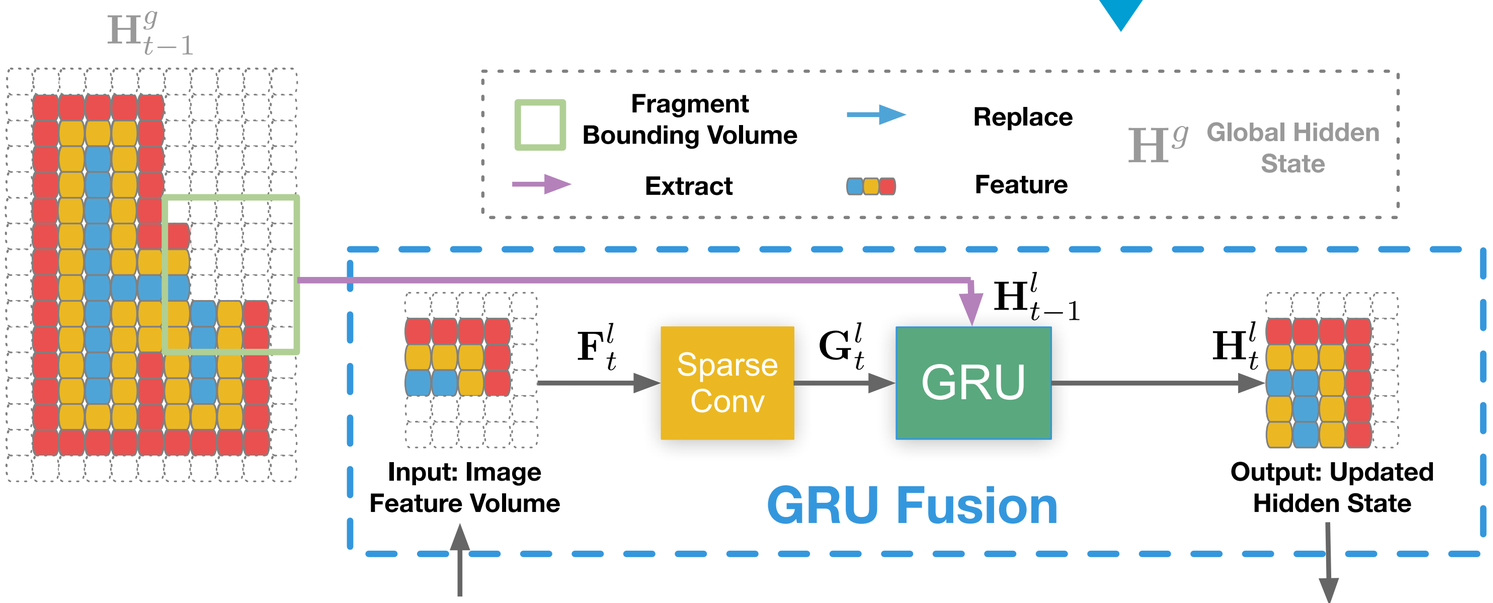


NeuralRecon

Joint reconstruction and fusion

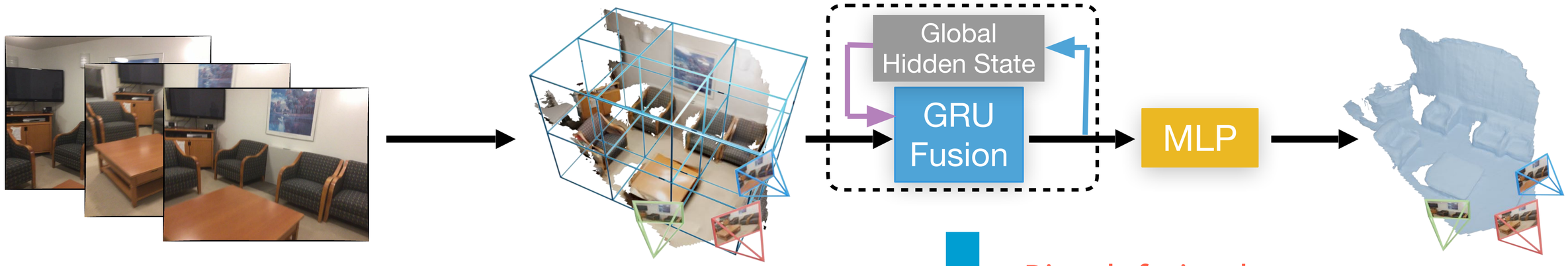


Directly fusing the features with GRU

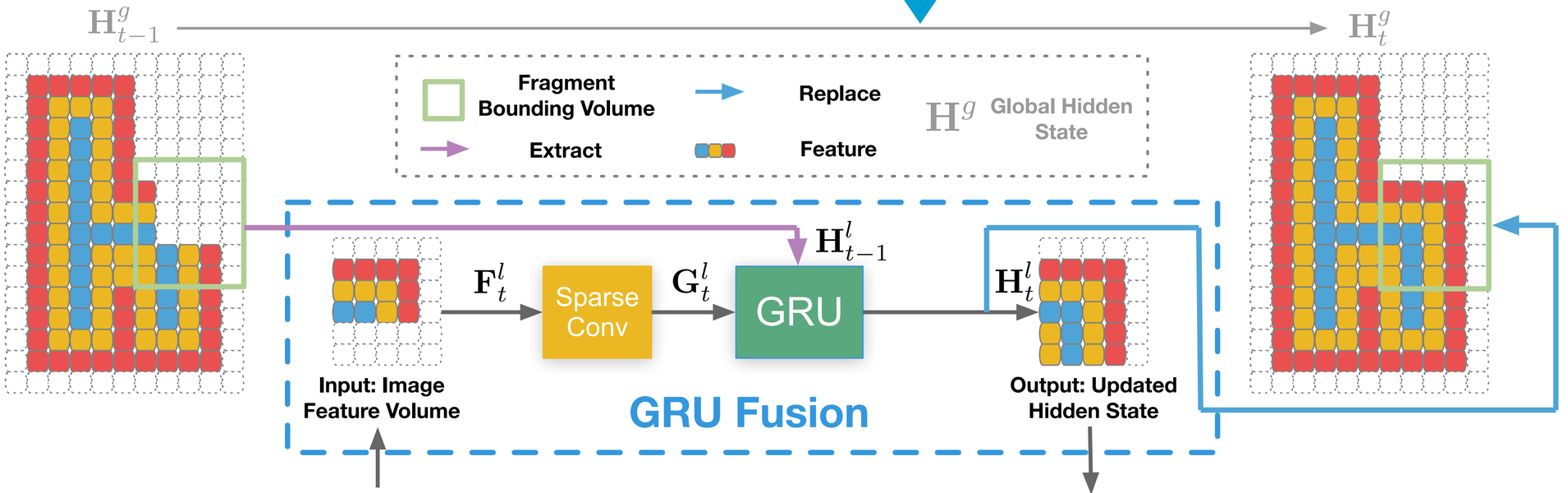


NeuralRecon

Joint reconstruction and fusion

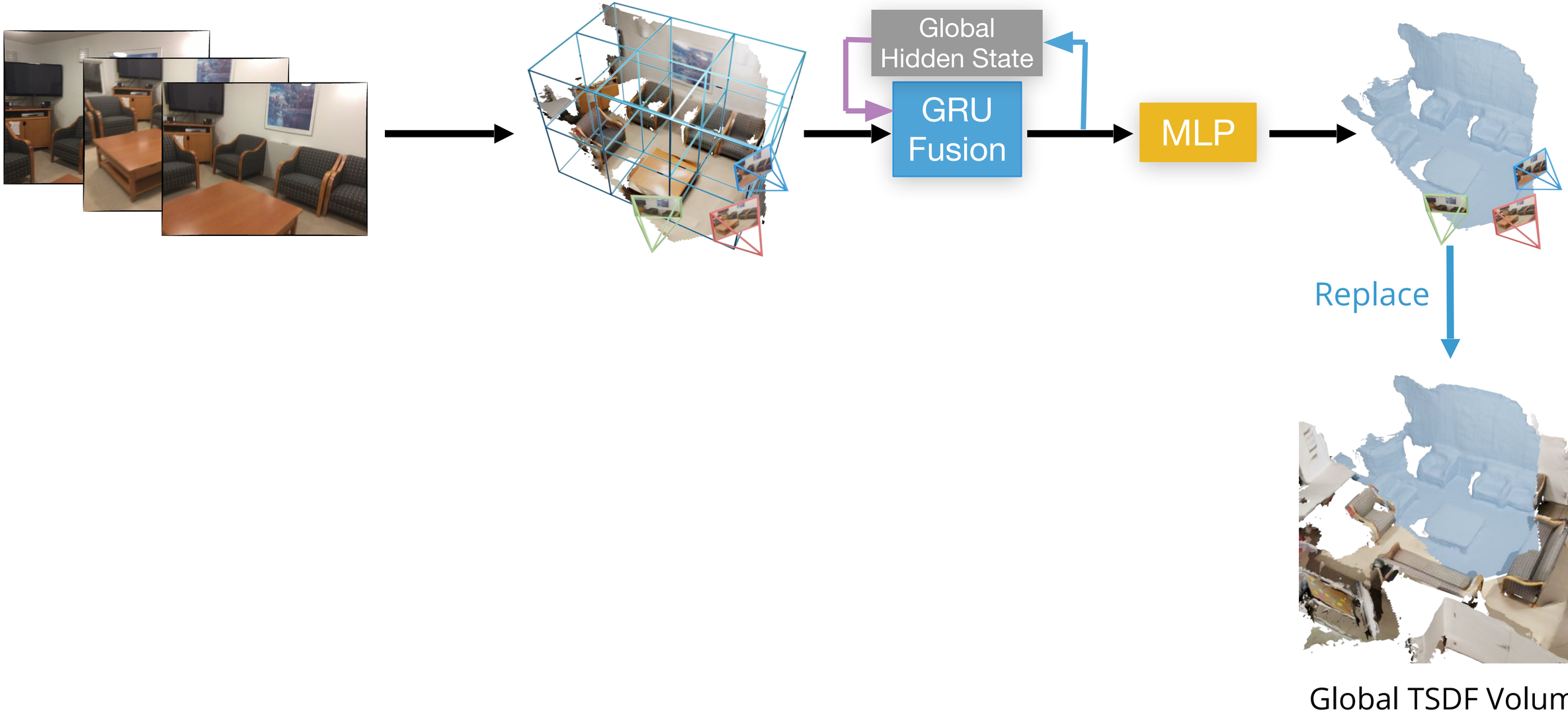


Directly fusing the features with GRU



NeuralRecon

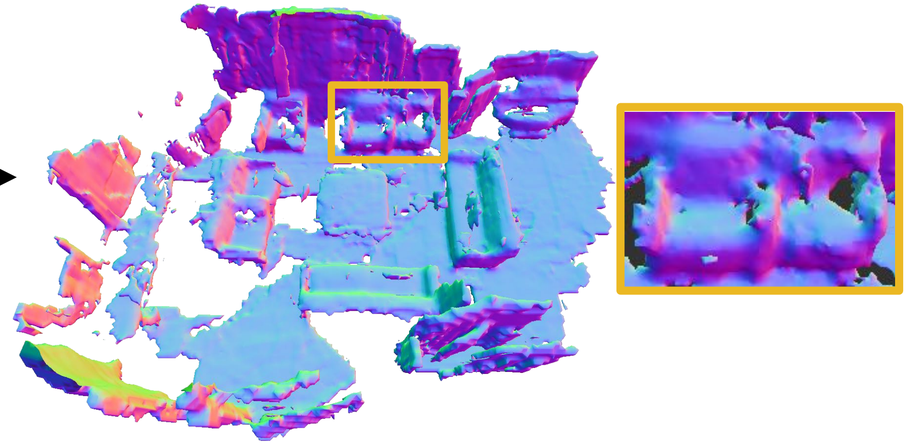
Joint reconstruction and fusion



NeuralRecon

Joint reconstruction and fusion

Fragment Reconstruction → TSDF Fusion →



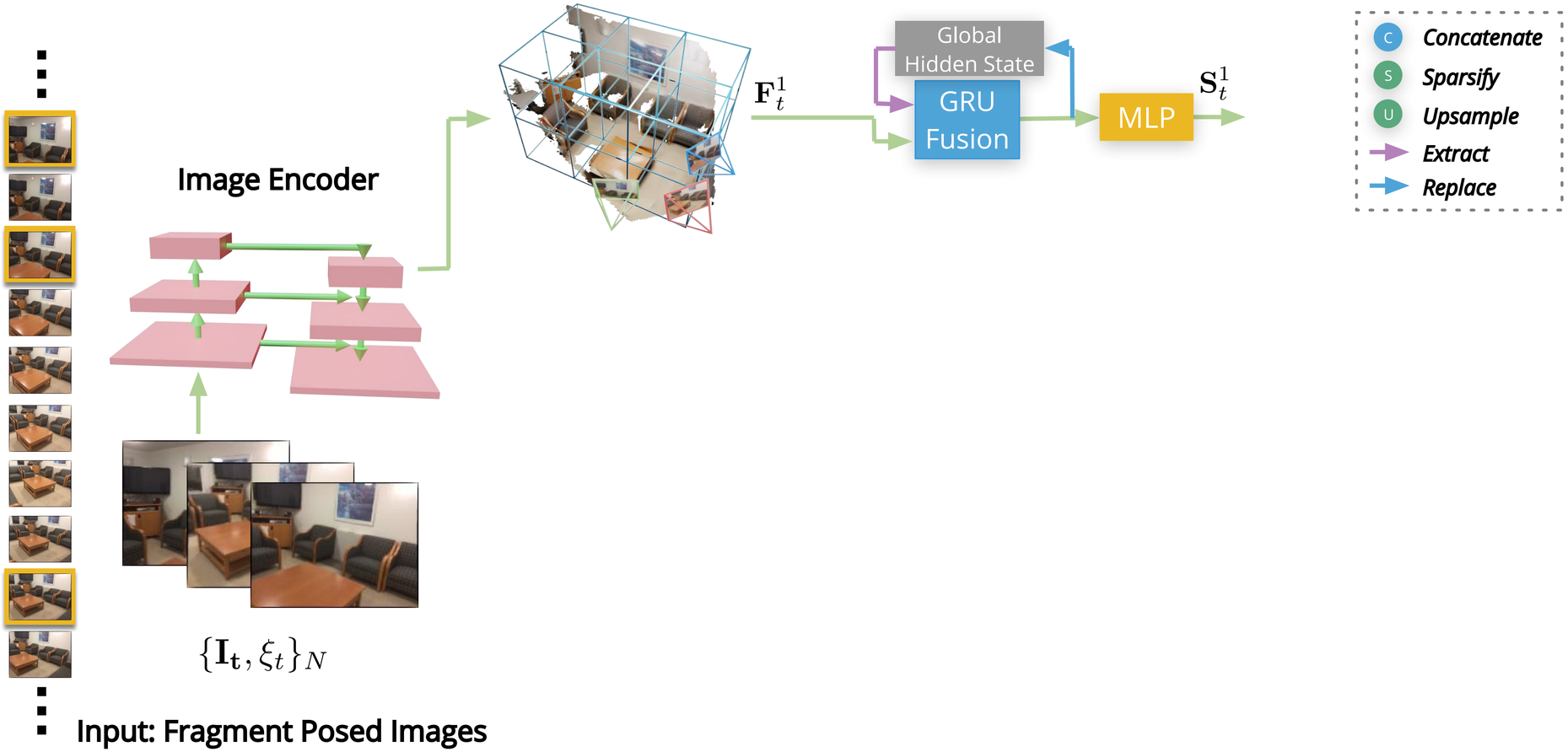
VS

😃 **Globally coherent**

Joint Reconstruction and Fusion →

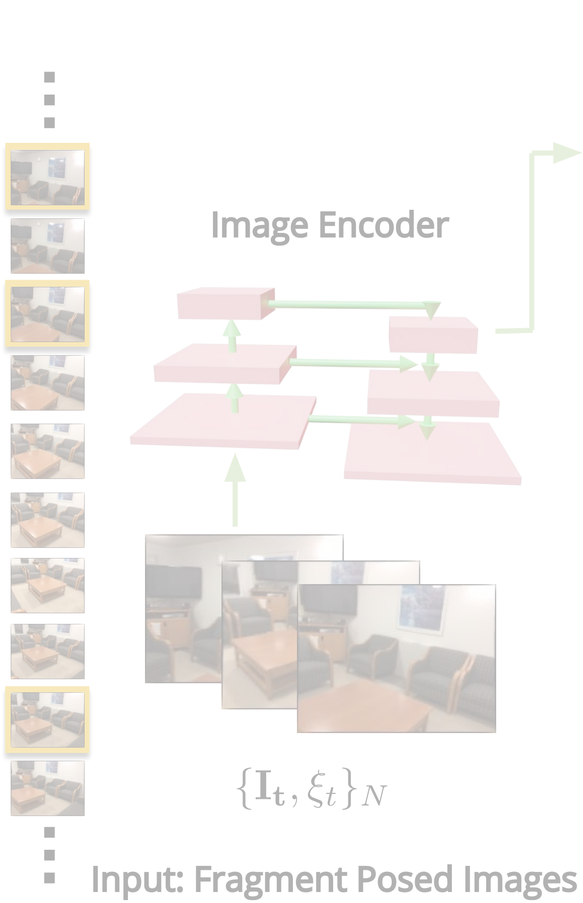
NeuralRecon

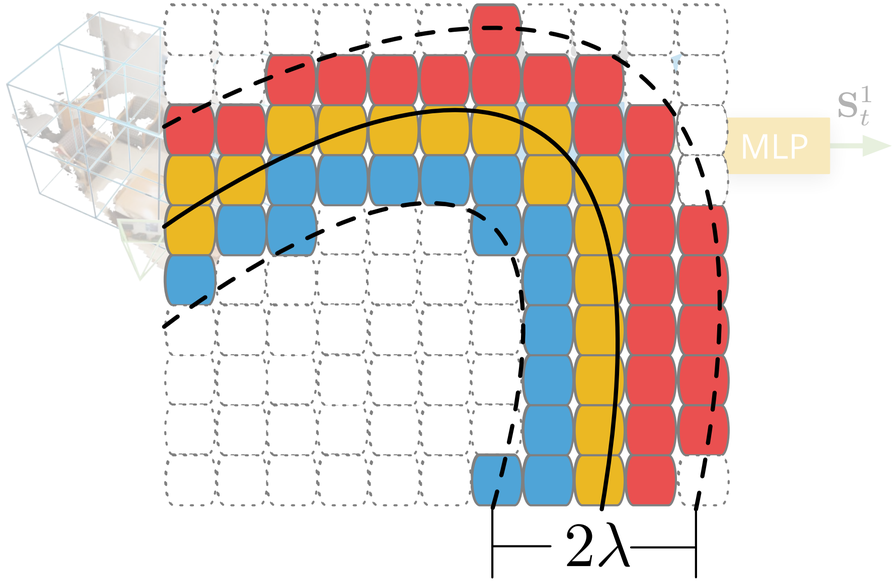
Coarse-to-fine architecture

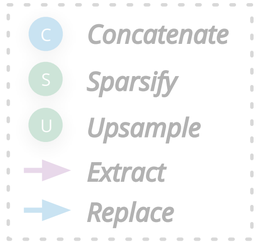


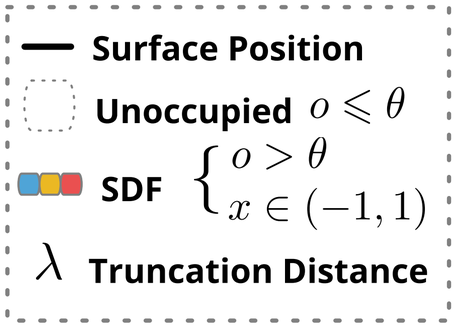
NeuralRecon

Coarse-to-fine architecture





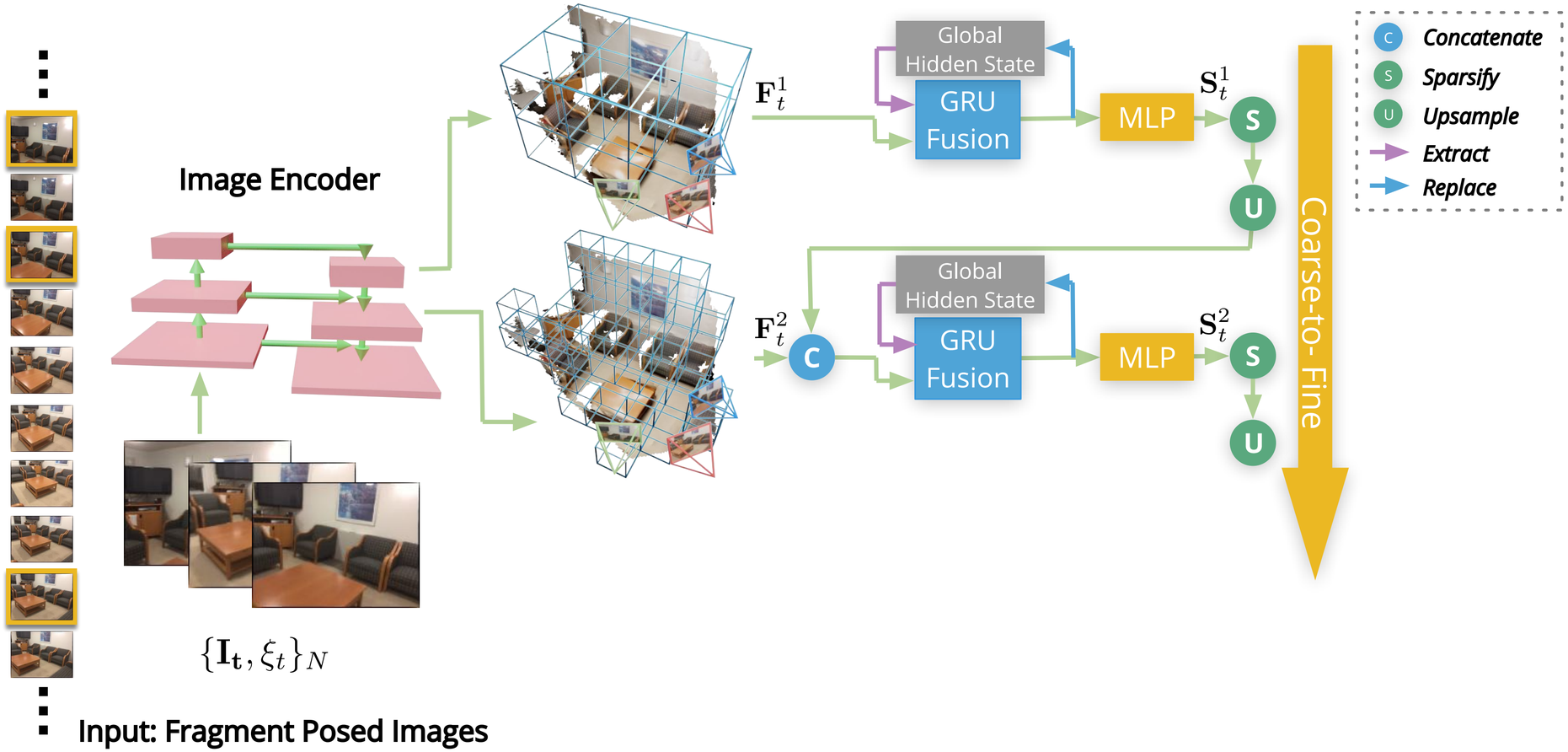




Output of *MLP* : **Occupancy Score** and **SDF**

NeuralRecon

Coarse-to-fine architecture

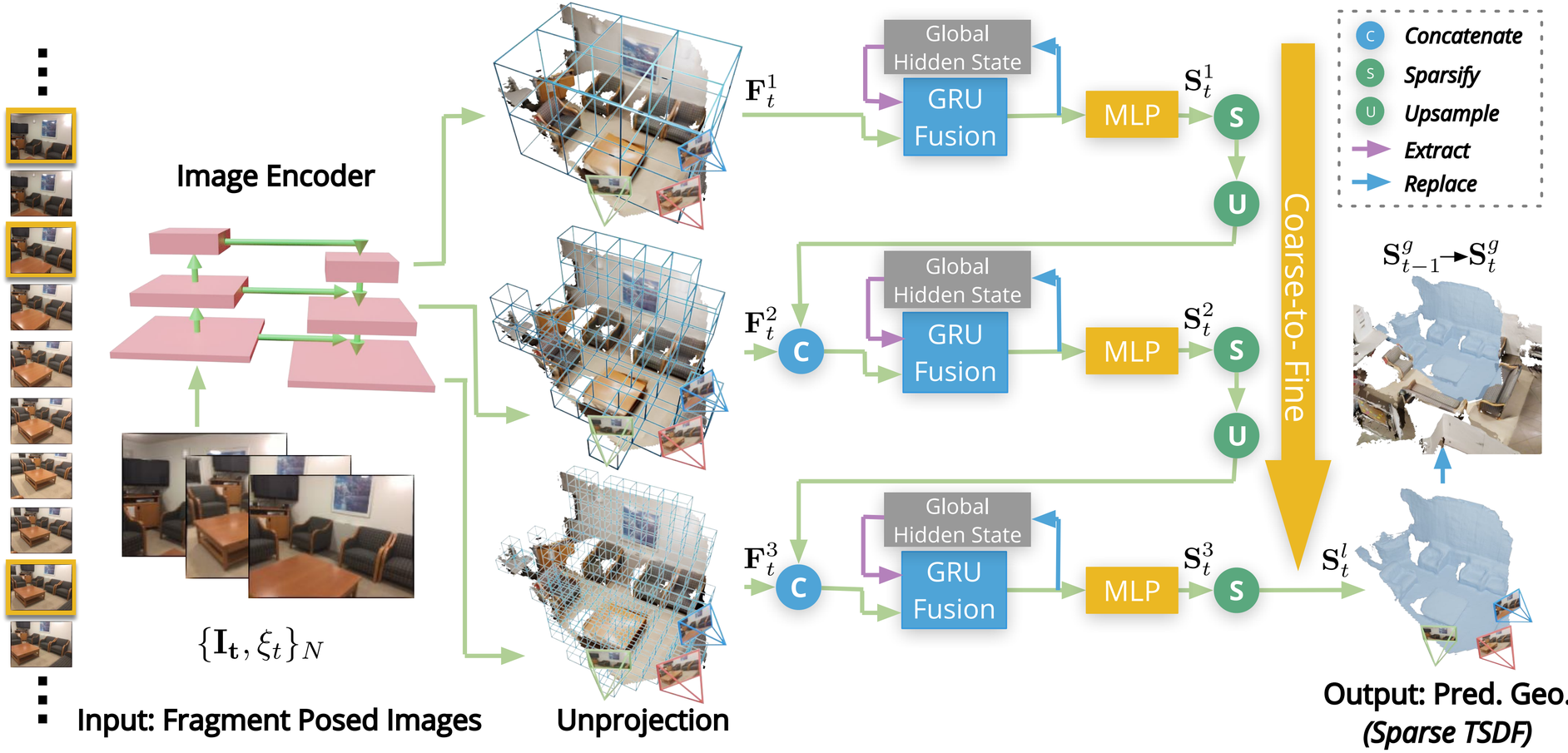


Output of *MLP* : **Occupancy Score** and **SDF**

S Filter by Occupancy Score > 0

NeuralRecon

Coarse-to-fine architecture

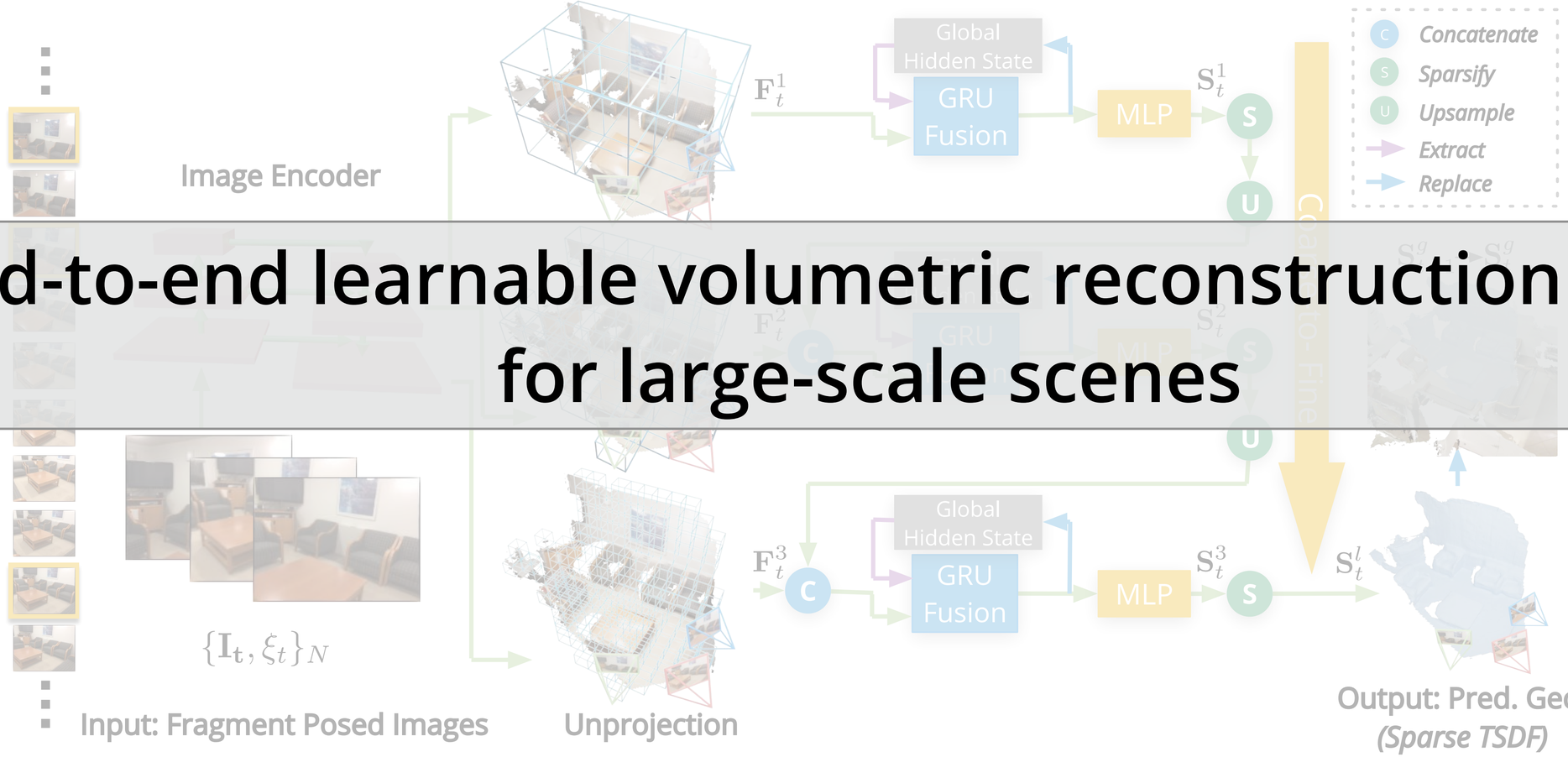


Output of *MLP* : **Occupancy Score** and **SDF**

S Filter by Occupancy Score > 0

NeuralRecon

Coarse-to-fine architecture



End-to-end learnable volumetric reconstruction system for large-scale scenes

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

S Filter by Occupancy Score > 0

NeuralRecon

Training

ScanNet dataset

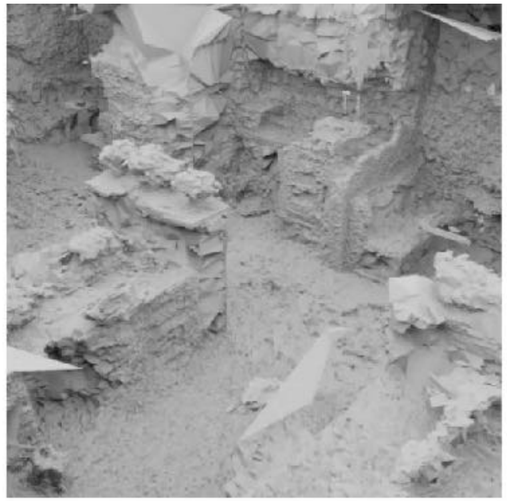
Contains 2.5M RGB images captured in 707 indoor scenes with ground-truth TSDF

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

Experiments

Qualitative results: office 1

Ours (30ms)

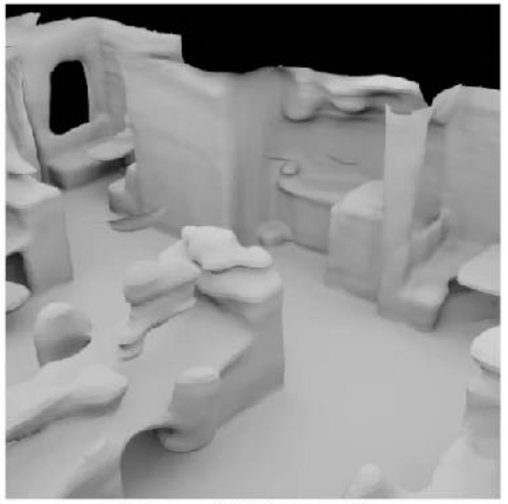


COLMAP (2076ms)

DeepV2D (347ms)

Ground Truth

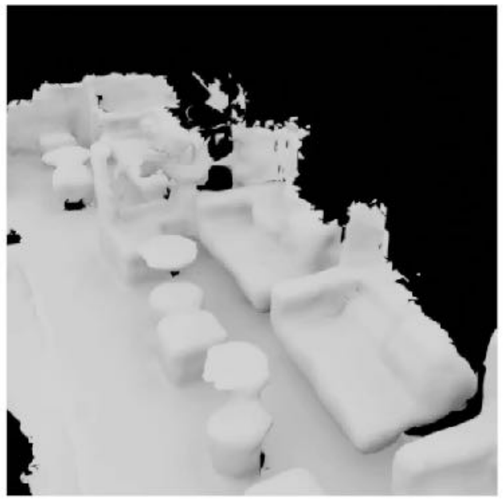
CNMNet (80ms)



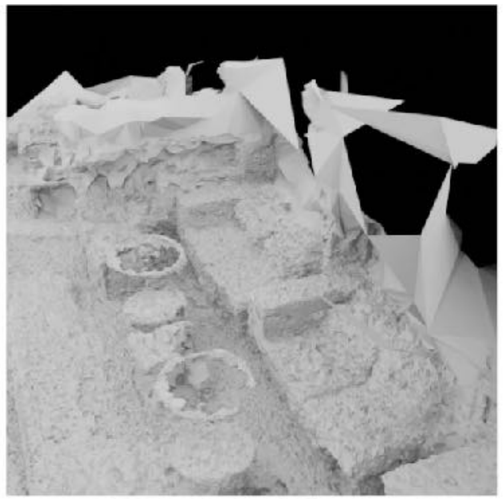
Atlas (292ms)

Experiments

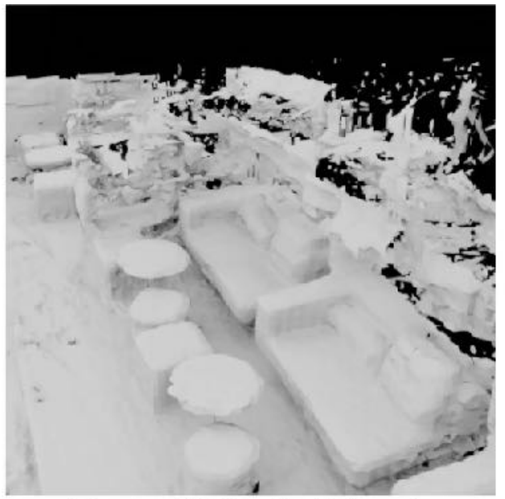
Qualitative results: office 2



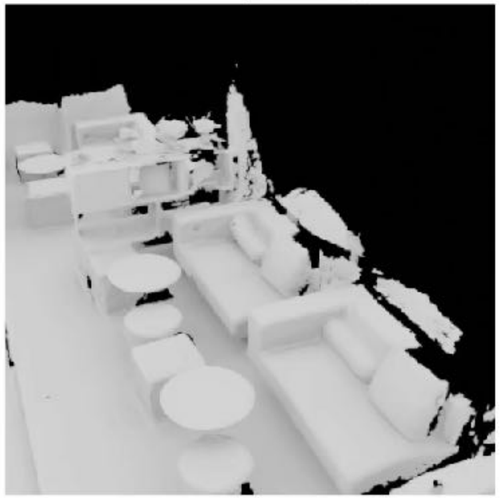
Ours (30ms)



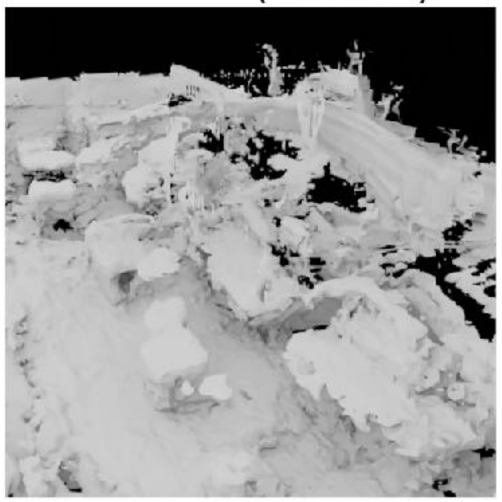
COLMAP (2076ms)



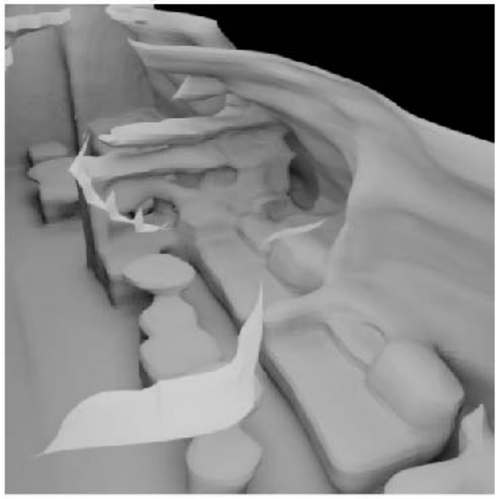
DeepV2D (347ms)



Ground Truth



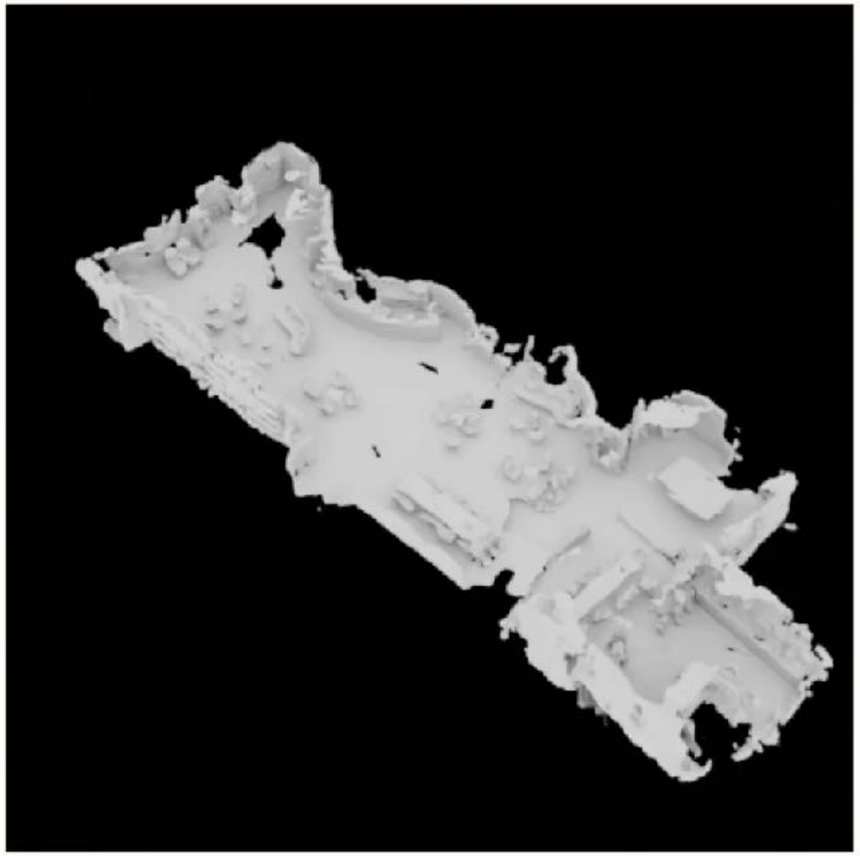
CNMNet (80ms)



Atlas (292ms)

Experiments

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



Ours

Max GPU Memory: 3.29GB

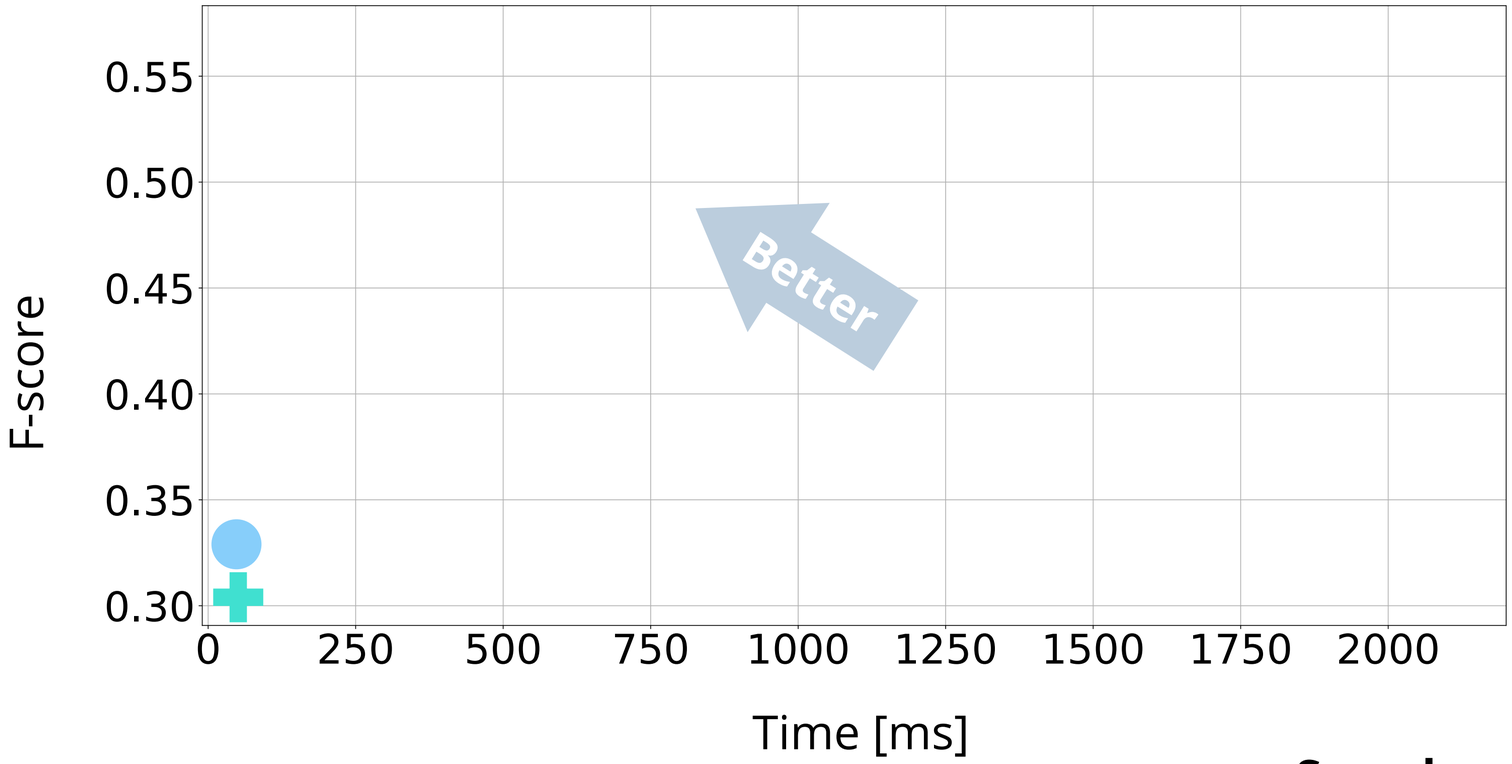
Atlas

Max GPU Memory: >24GB (OOM)

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

Experiments

Quantitive results



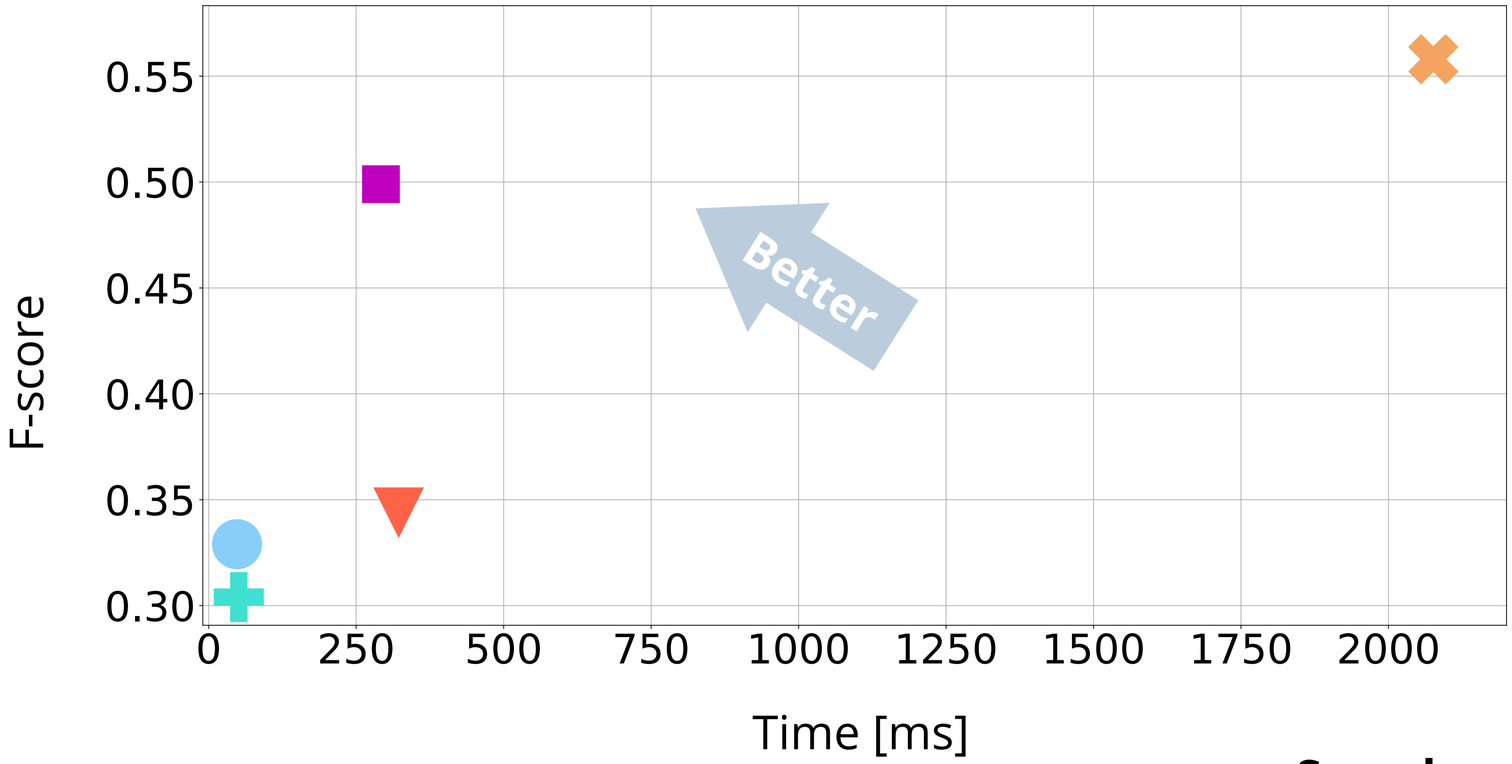
Real-time methods: MVDepthNet GPMVS

Speed 😃

Accuracy 😕

Experiments

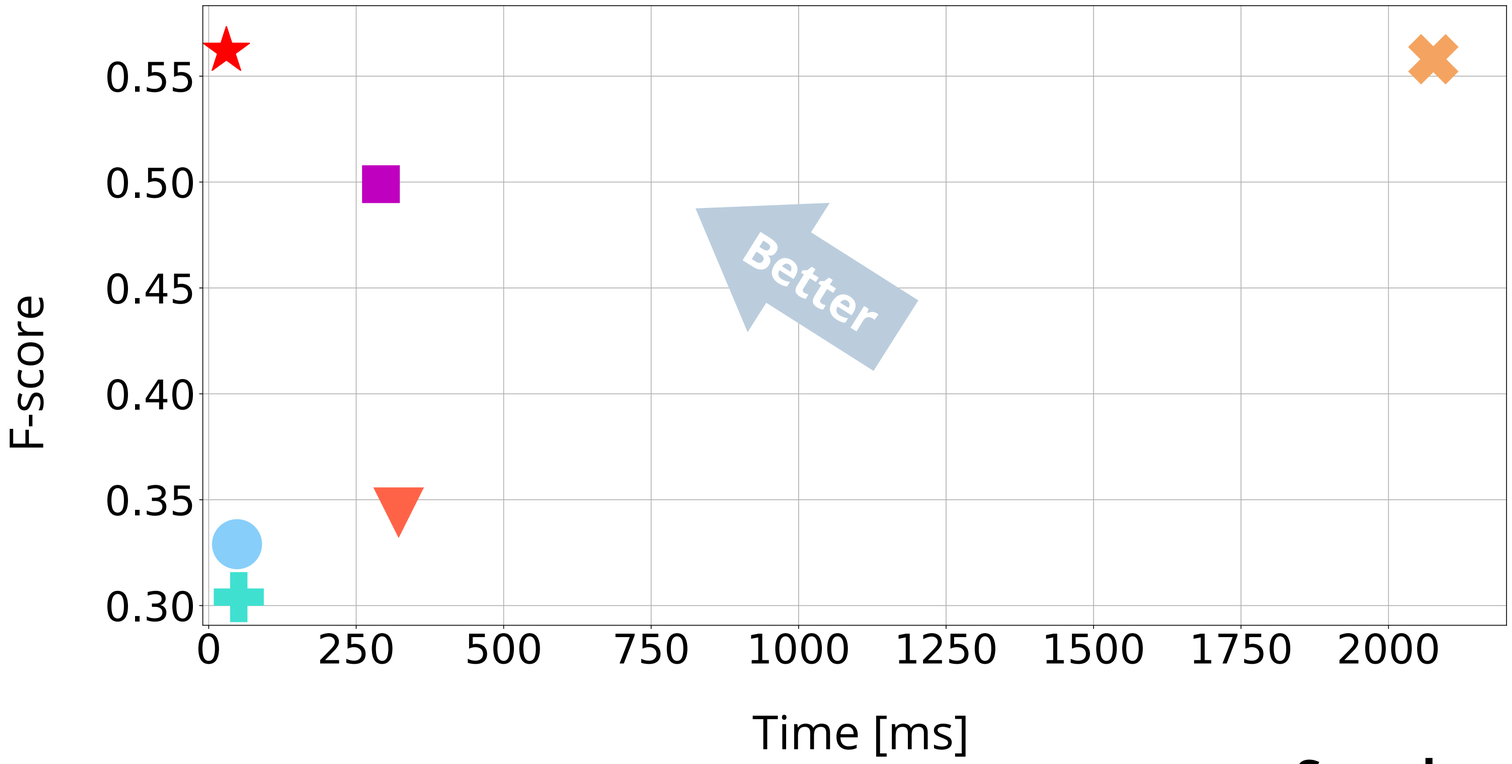
Quantitive results



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

Experiments

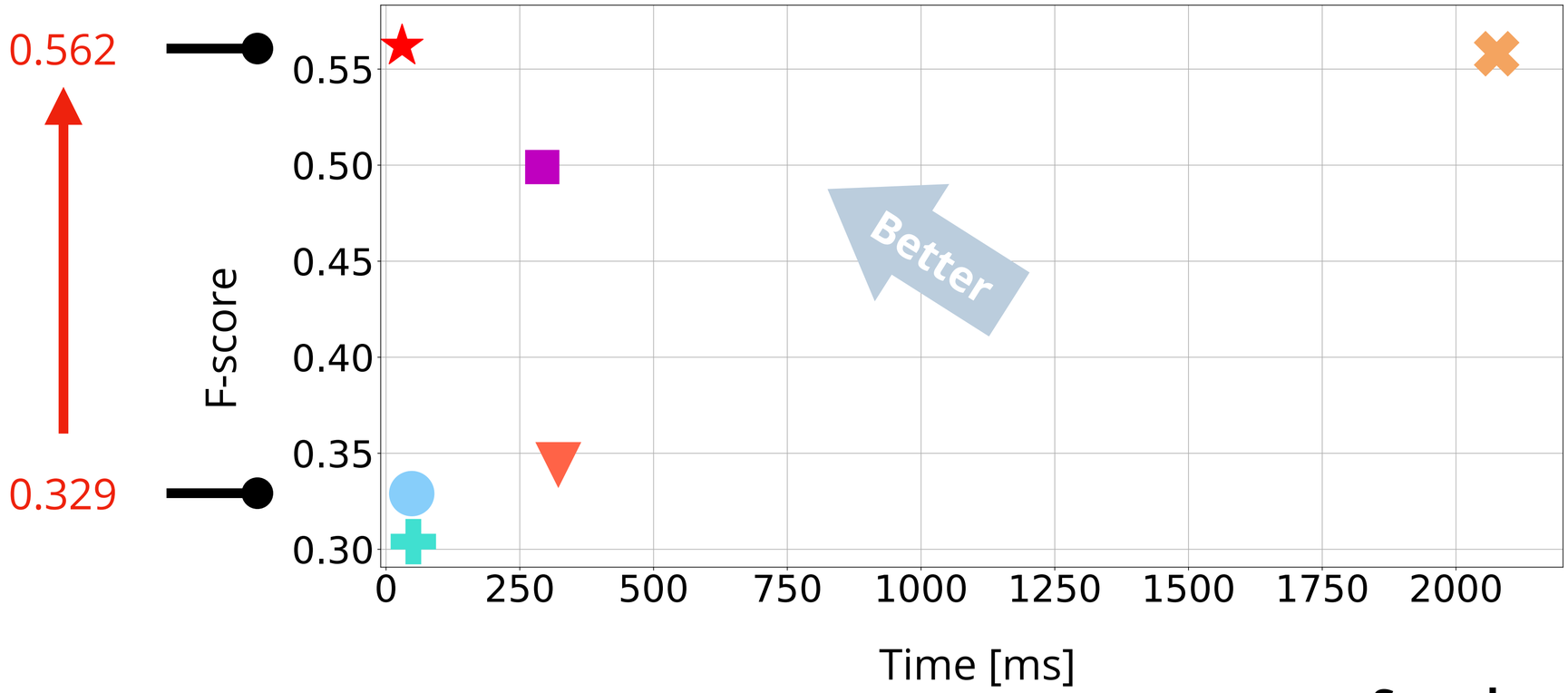
Quantitive results



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

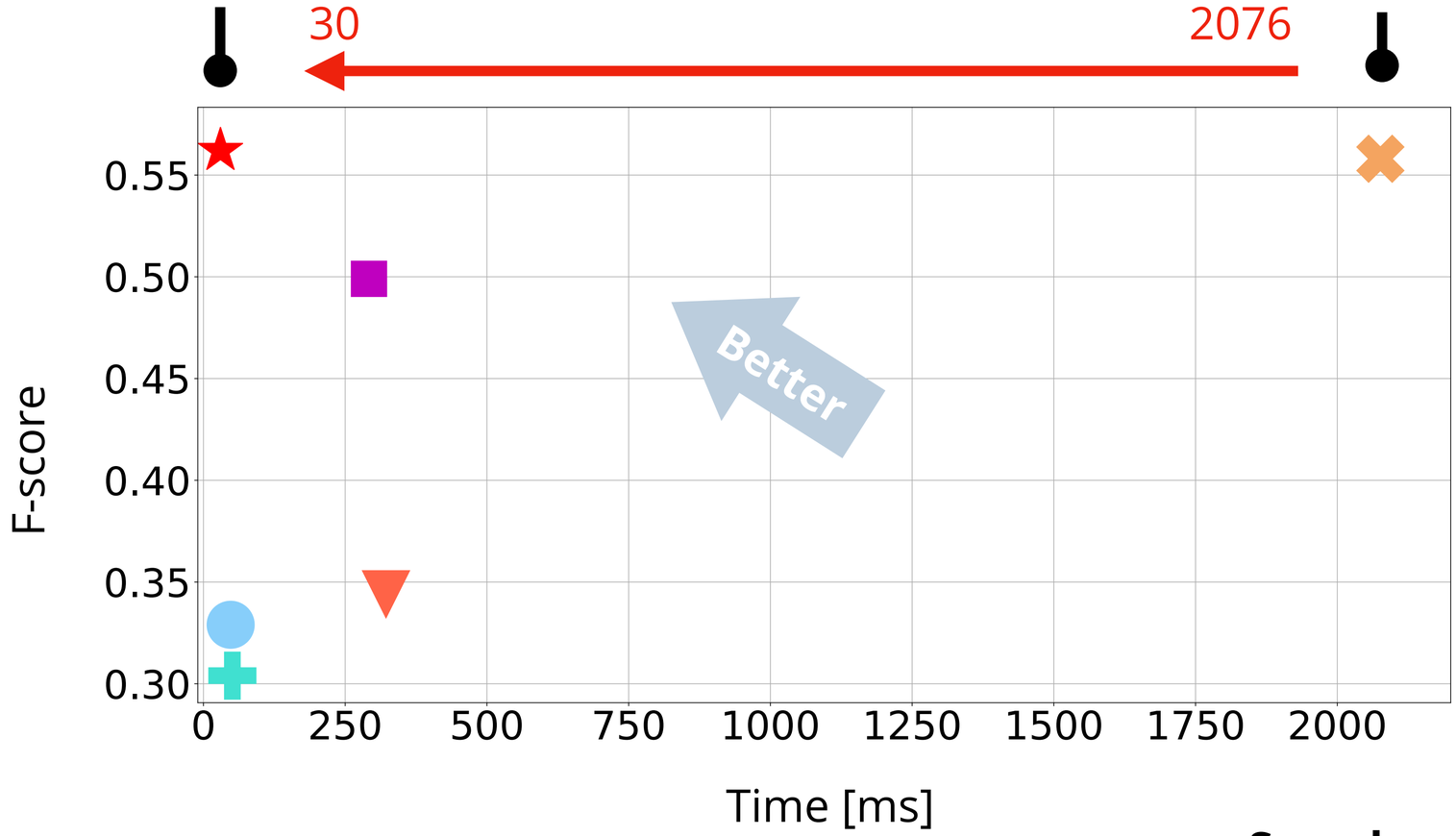
Experiments

Quantitive results



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

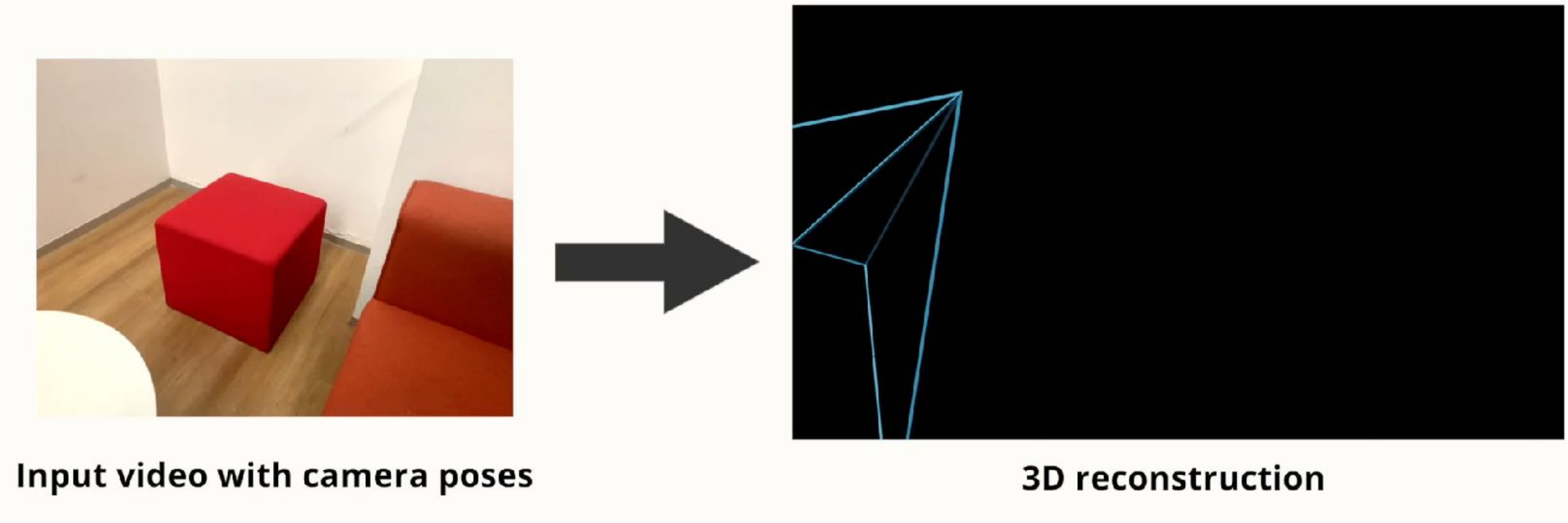
Experiments



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

Demo

Indoor scene at our office

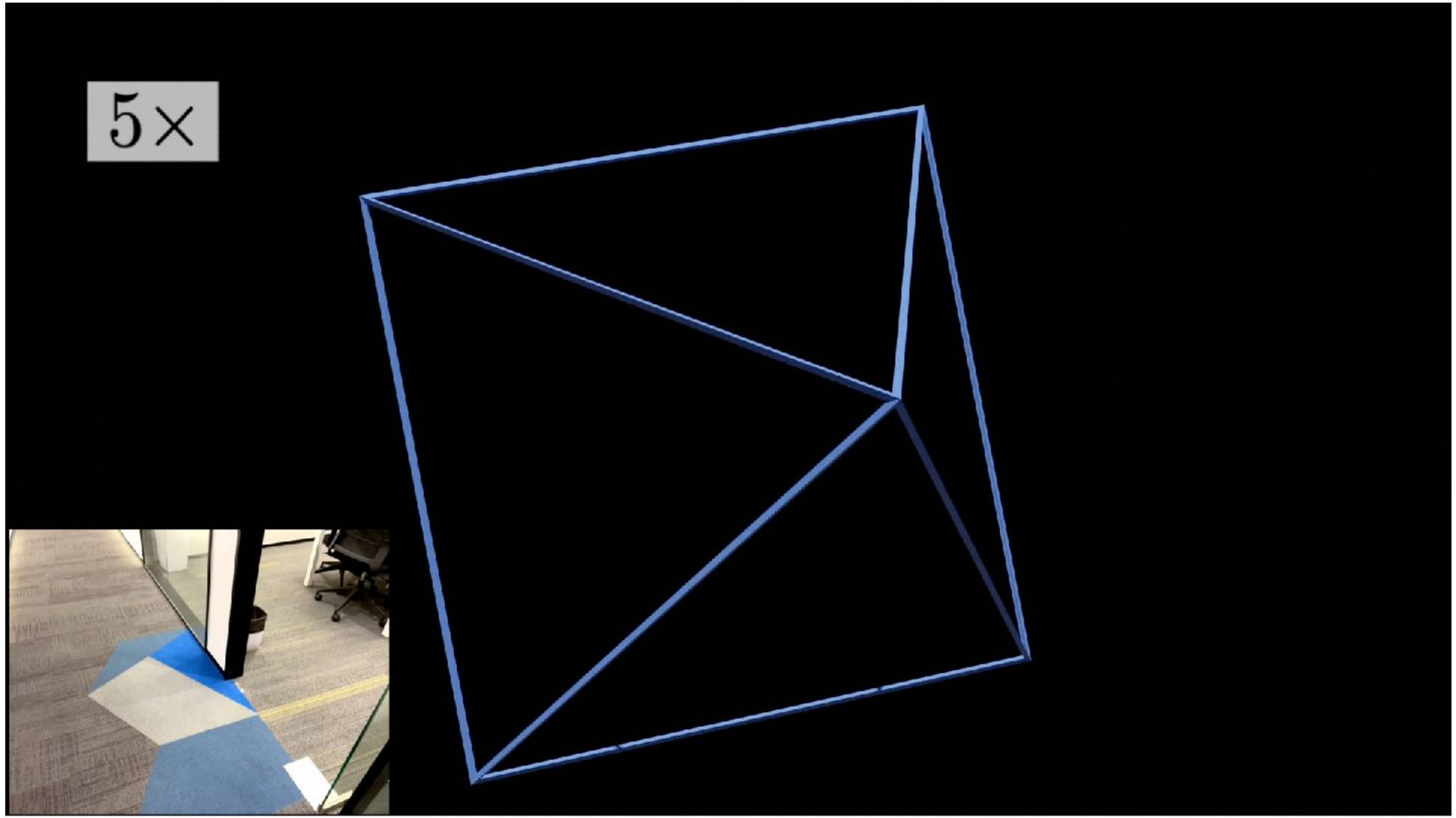


Input video with camera poses

3D reconstruction

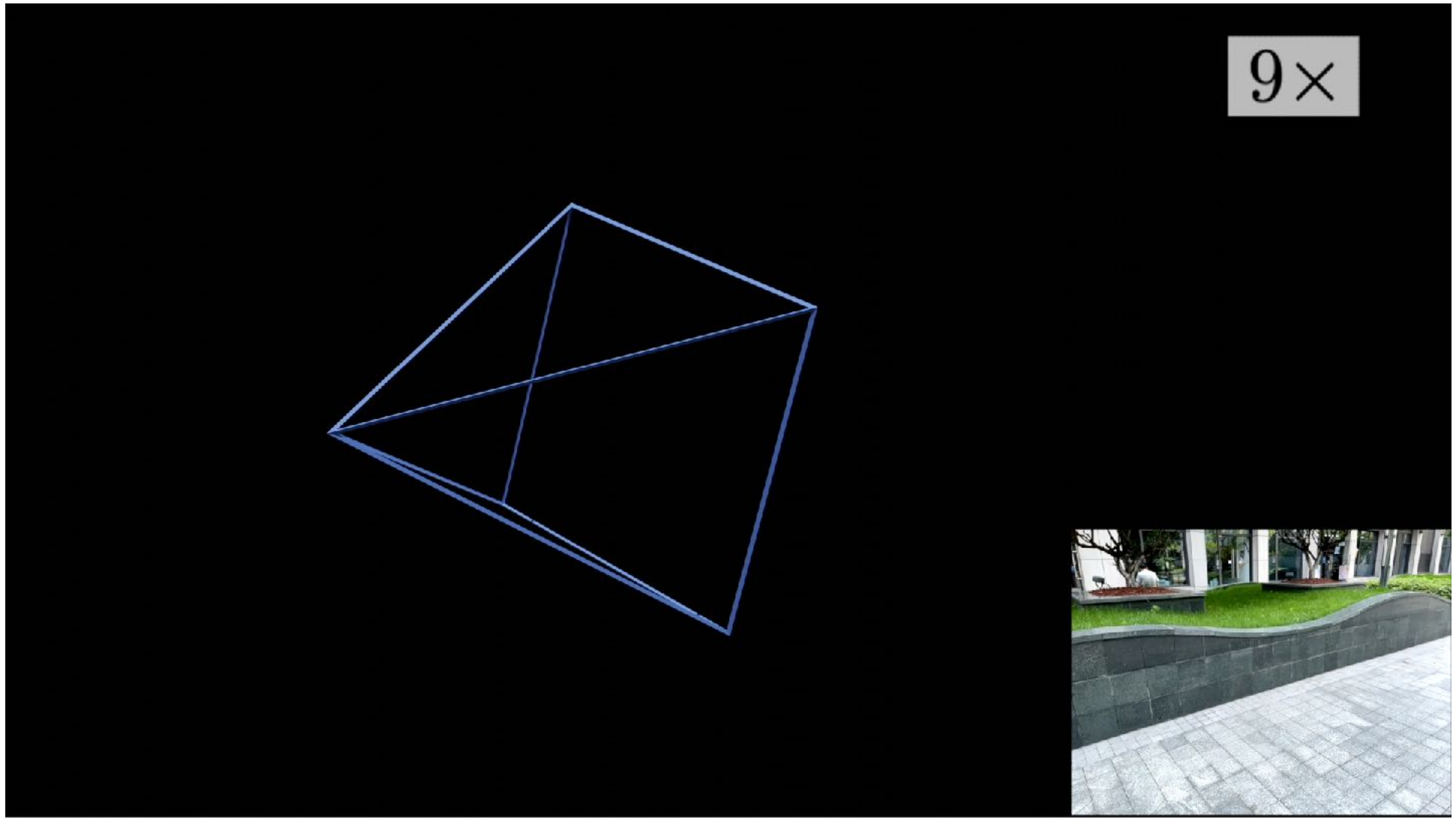
Demo

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

Conclusion and Discussion

- **Learning-based 3D reconstruction and localization**
  - Determine what's really necessary to be learned
    (e.g. features for matching, surface prior for reconstruction and fusion. BA, PnP, RANSAC are fine and should be kept.)
  - Design end-to-end learnable systems with differentiable modules
    (e.g. GRU-based TSDF reconstruction and fusion in NeuralRecon, differentiable matching in LoFTR, differentiable and learnable optimization, etc.)

Conclusion and Discussion

- **Learning-based 3D reconstruction and localization**
  - Determine what's really necessary to be learned
    (e.g. features for matching, surface prior for reconstruction and fusion. BA, PnP, RANSAC are fine and should be kept.)
  - Design end-to-end learnable systems with differentiable modules
    (e.g. GRU-based TSDF reconstruction and fusion in NeuralRecon, differentiable matching in LoFTR, differentiable and learnable optimization, etc.)
- **Future 3D perception systems?**
  - Learnable and embodied (e.g. semantics, context-dependent)
  - Self-supervised from observations and interactions
    (e.g. with differentiable rendering, simulators)
  - Hardcoded with well-established knowledges as prior
    (e.g. multi-view geometry, laws of physics)

Acknowledgements

Xiaowei Zhou

Yiming Xie

Zehong Shen

Yu’ang Wang

Linghao Chen

Thanks for watching!
Q&A