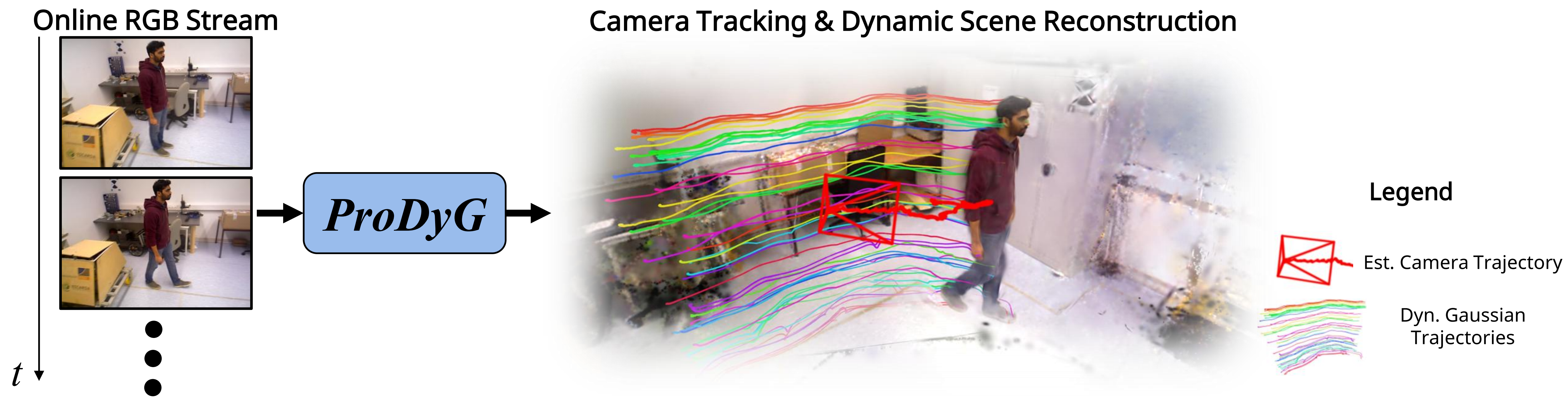
ETH zürich

ProDyG: **Pro**gressive **Dy**namic Scene Reconstruction via **G**aussian Splatting from Monocular Videos

Shi Chen[1], Erik Sandström[2], Sandro Lombardi, Siyuan Li[1], Martin R. Oswald[3]
[1]ETH Zürich, [2]Google, [3]University of Amsterdam

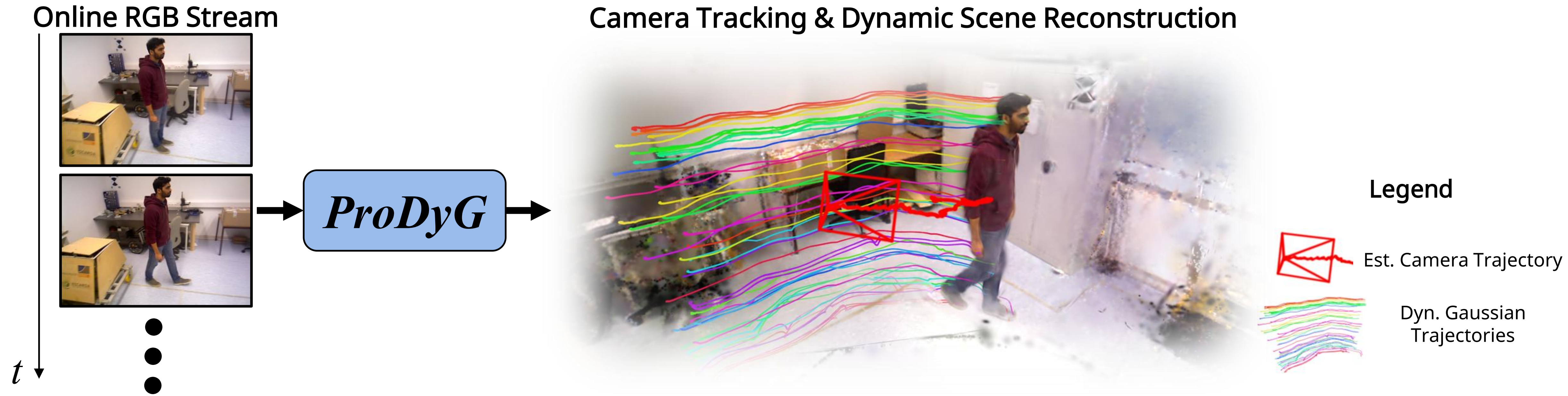


Motivation: What is needed for practical dynamic scene reconstruction?

- **Monocular** input
- **Online** operation
- **Robust camera tracking** against dynamic distractors
- **Expressive** scene representation
- **Temporal consistency**
- **NVS** of **dynamic** regions
- **RGB-only**

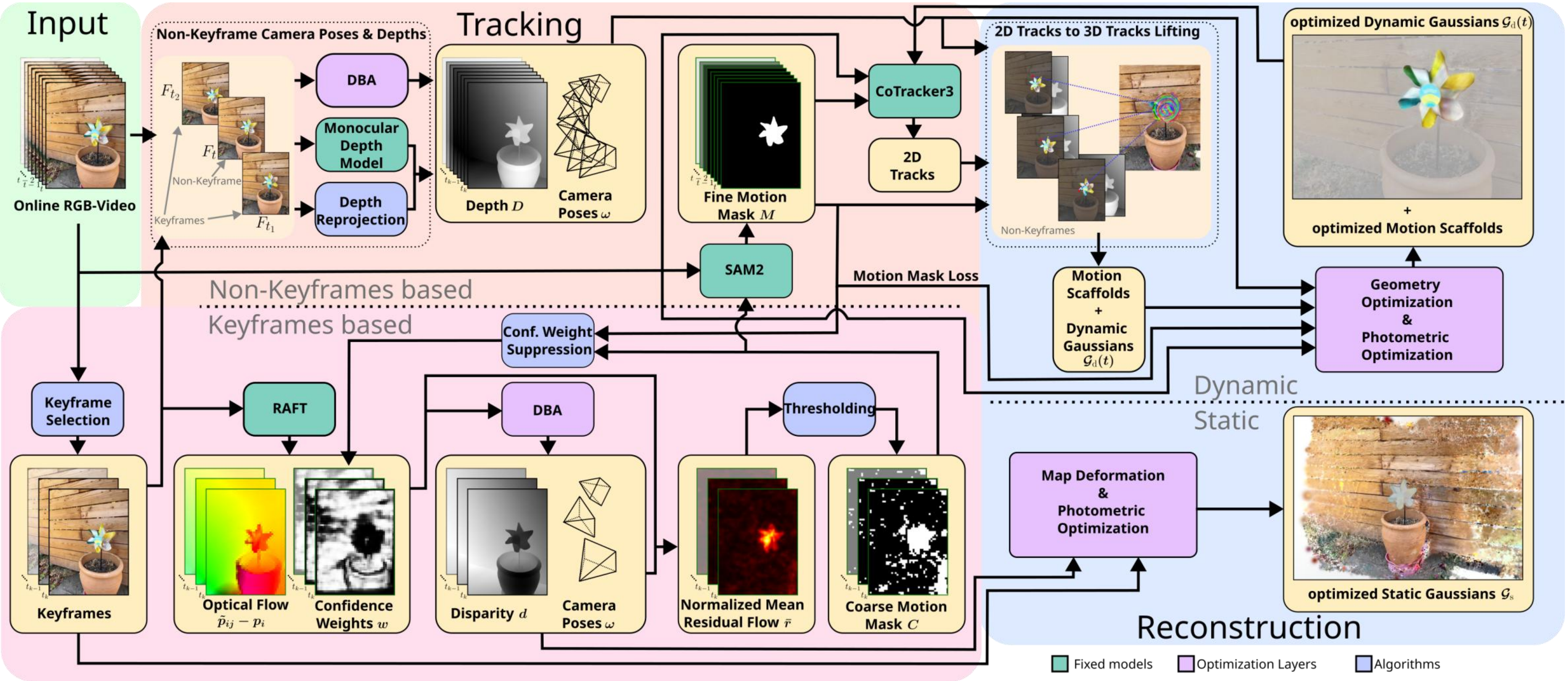
ProDyG ticks all the boxes!

- ✓ Monocular input
- ✓ Online operation
- ✓ Robust camera tracking against dynamic distractors
- ✓ Expressive scene representation
- ✓ Temporal consistency
- ✓ NVS of dynamic regions
- ✓ RGB-only



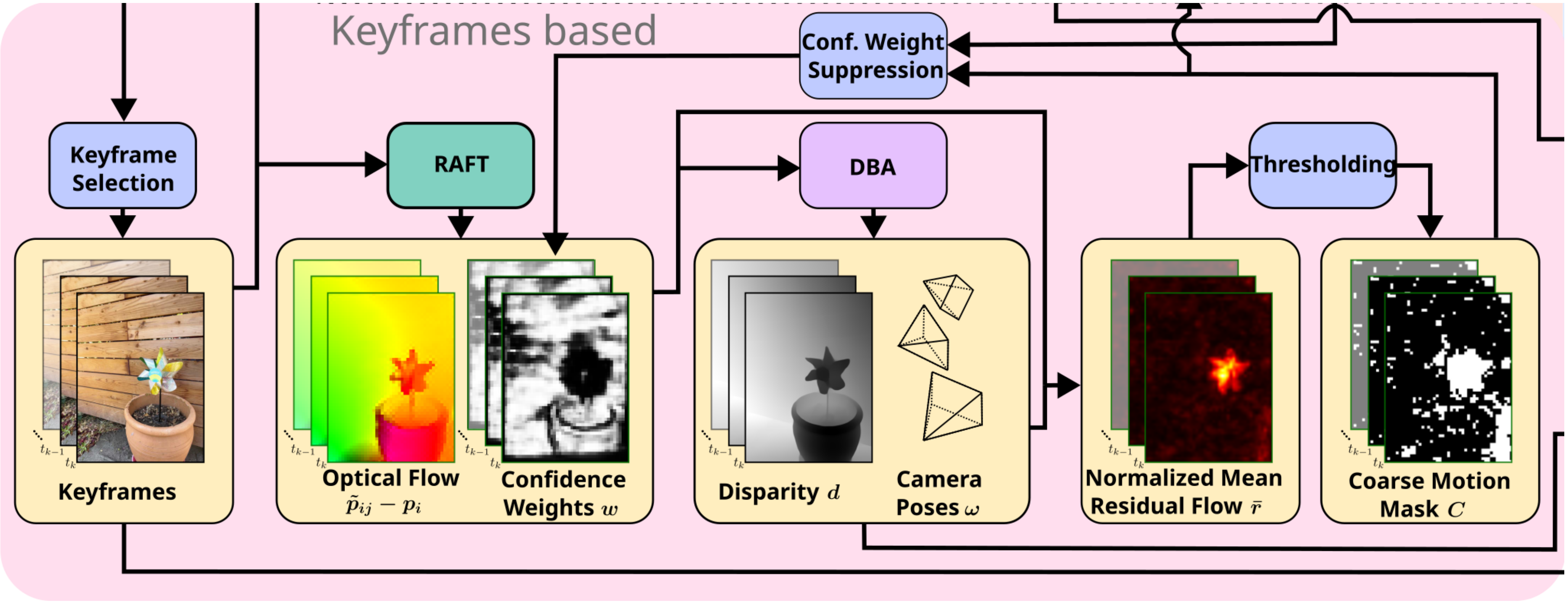
Method Overview

- Motion-Agnostic Online Camera Tracking
- Progressive Dynamic Scene Reconstruction



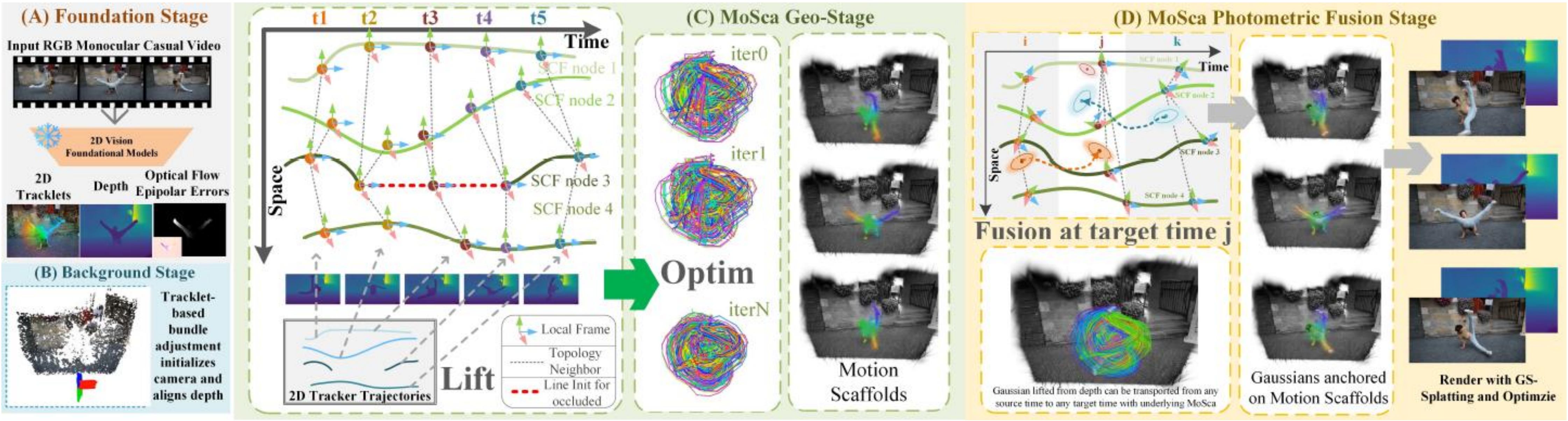
Motion-Agnostic Online Camera Tracking

- Operates over a **factor graph**
- Iteratively refine **coarse motion mask** based on **residual flow** = optical flow – camera-induced flow
- **Suppress confidence** in potentially dynamic regions



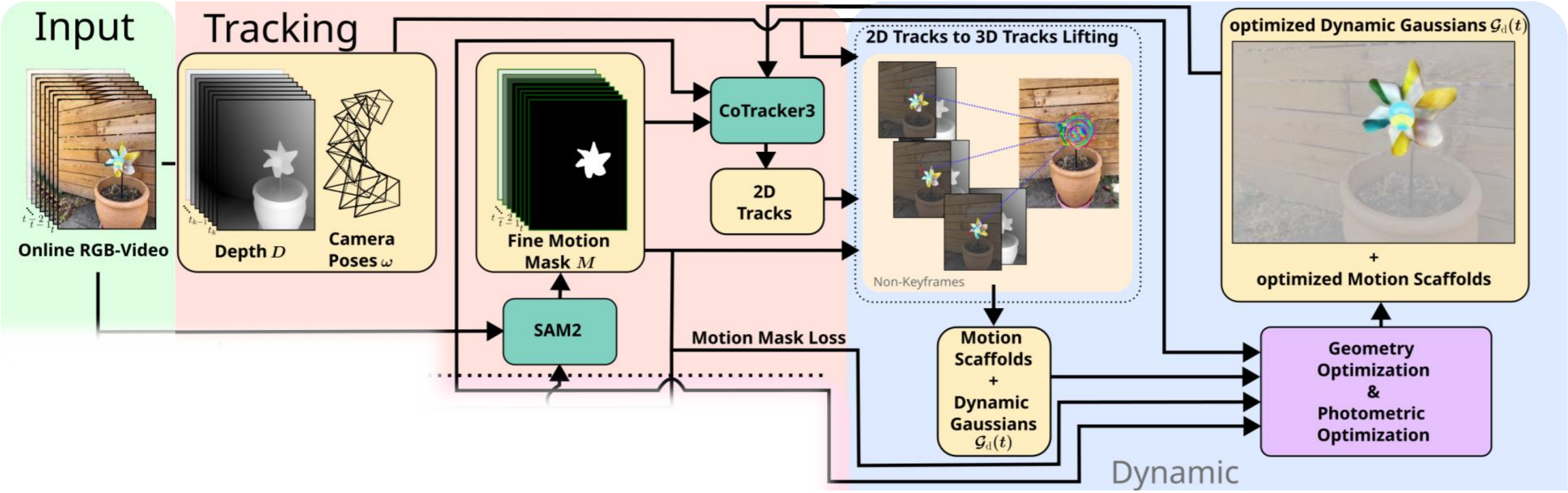
Motion Scaffolds (MoSca) [Lei 2024]

- Low-parametric motion field representation
- Composed of nodes with time-dependent positions & rotations and edges representing topology
- Motion of dynamic Gaussians computed by interpolating MoSca node motion using Dual Quaternion Blending (DQB)



Progressive Construction of Motion Scaffolds

- We extend **offline** MoSca to **online**
  - ➢ Extend 2D tracks into new frames
  - ➢ Identify newly observed pixels
  - ➢ Add new 2D tracks
  - ➢ Lift 2D tracks into 3D, and warp them backward in time using Motion Scaffolds
  - ➢ Geometry & photometric optimization



Tracking: Quantitative Results (ATE RMSE [cm])

Method	Type	Bonn RGB-D Dynamic Dataset [47]					TUM RGB-D Dataset [62]				
		Ball	Ball2	Pers	Pers2	Avg.	f3/ws	f3/wx	f3/wr	f3/whs	Avg.
RGB-D Input											
ORB-SLAM2 [44]	S	6.5	23.0	6.9	7.9	11.1	40.8	72.2	80.5	72.3	66.45
NICE-SLAM [90]	S	24.4	20.2	24.5	53.6	30.7	79.8	86.5	244.0	152.0	140.57
ReFusion [48]	R	17.5	25.4	28.9	46.3	29.5	1.7	9.9	40.6	10.4	15.7
DynaSLAM (N+G) [4]	R	3.0	2.9	6.1	7.8	5.0	0.6	1.5	3.5	2.5	2.03
DG-SLAM [76]	R	3.7	4.1	4.5	6.9	4.8	0.6	1.6	4.3	-	-
RoDyn-SLAM [19]	R	7.9	11.5	14.5	13.8	11.9	1.7	8.3	-	5.6	-
DDN-SLAM (RGB-D) [33]	R	**1.8**	4.1	4.3	3.8	3.5	1.0	1.4	3.9	2.3	2.15
RGB Input											
DSO [10]	S	7.3	21.8	30.6	26.5	21.6	1.5	12.9	13.8	40.7	17.23
DROID-SLAM [66]	S	7.5	4.1	4.3	5.4	5.3	1.2	1.6	4.0	2.2	2.25
MonoGS [42]	S	15.3	17.3	26.4	35.2	23.6	1.1	21.5	17.4	44.2	21.05
Splat-SLAM [54]	S	8.8	3.0	4.9	25.8	10.6	2.3	1.3	3.9	2.2	2.43
DDN-SLAM (RGB) [33]	R	-	-	-	-	-	2.5	2.8	8.9	4.1	4.58
MegaSaM [37]	R	3.7	2.6	4.1	4.0	3.6	0.6	1.5	**2.6**	1.8	**1.63**
WildGS-SLAM [89]	R	2.7	**2.4**	3.6	3.1	**2.94**	**0.4**	1.3	3.3	**1.6**	**1.63**
DynaMoN (MS) [55]	D	6.8	3.8	**2.4**	3.5	4.1	1.4	1.4	3.9	2.0	2.18
DynaMoN (MS&SS) [55]	D	2.8	2.7	14.8	**2.2**	5.6	0.7	1.4	3.9	1.9	1.98
D4DGS-SLAM* [65]	D	3.6	3.9	4.5	5.2	4.3	-	-	-	-	-
4D-GS SLAM* [35]	D	2.4	3.7	8.9	9.4	6.1	0.5	2.1	**2.6**	-	-
ProDyG (Ours)	D	2.7	2.6	4.9	2.9	3.29	1.6	**1.2**	3.0	1.7	1.89

Tracking: Quantitative Results (ATE RMSE [cm])

- Type S: Static scenes

Method	Type	Bonn RGB-D Dynamic Dataset [47]					TUM RGB-D Dataset [62]				
		Ball	Ball2	Pers	Pers2	Avg.	f3/ws	f3/wx	f3/wr	f3/whs	Avg.
RGB-D Input											
ORB-SLAM2 [44]	S	6.5	23.0	6.9	7.9	11.1	40.8	72.2	80.5	72.3	66.45
NICE-SLAM [90]	S	24.4	20.2	24.5	53.6	30.7	79.8	86.5	244.0	152.0	140.57
ReFusion [48]	R	17.5	25.4	28.9	46.3	29.5	1.7	9.9	40.6	10.4	15.7
DynaSLAM (N+G) [4]	R	3.0	2.9	6.1	7.8	5.0	0.6	1.5	3.5	2.5	2.03
DG-SLAM [76]	R	3.7	4.1	4.5	6.9	4.8	0.6	1.6	4.3	-	-
RoDyn-SLAM [19]	R	7.9	11.5	14.5	13.8	11.9	1.7	8.3	-	5.6	-
DDN-SLAM (RGB-D) [33]	R	**1.8**	4.1	4.3	3.8	3.5	1.0	1.4	3.9	2.3	2.15
RGB Input											
DSO [10]	S	7.3	21.8	30.6	26.5	21.6	1.5	12.9	13.8	40.7	17.23
DROID-SLAM [66]	S	7.5	4.1	4.3	5.4	5.3	1.2	1.6	4.0	2.2	2.25
MonoGS [42]	S	15.3	17.3	26.4	35.2	23.6	1.1	21.5	17.4	44.2	21.05
Splat-SLAM [54]	S	8.8	3.0	4.9	25.8	10.6	2.3	1.3	3.9	2.2	2.43
DDN-SLAM (RGB) [33]	R	-	-	-	-	-	2.5	2.8	8.9	4.1	4.58
MegaSaM [37]	R	3.7	2.6	4.1	4.0	3.6	0.6	1.5	**2.6**	1.8	**1.63**
WildGS-SLAM [89]	R	2.7	**2.4**	3.6	3.1	**2.94**	**0.4**	1.3	3.3	**1.6**	**1.63**
DynaMoN (MS) [55]	D	6.8	3.8	**2.4**	3.5	4.1	1.4	1.4	3.9	2.0	2.18
DynaMoN (MS&SS) [55]	D	2.8	2.7	14.8	**2.2**	5.6	0.7	1.4	3.9	1.9	1.98
D4DGS-SLAM* [65]	D	3.6	3.9	4.5	5.2	4.3	-	-	-	-	-
4D-GS SLAM* [35]	D	2.4	3.7	8.9	9.4	6.1	0.5	2.1	**2.6**	-	-
ProDyG (Ours)	D	2.7	2.6	4.9	2.9	3.29	1.6	**1.2**	3.0	1.7	1.89

Tracking: Quantitative Results (ATE RMSE [cm])

- Type S: Static scenes, Type R: Robust against dynamics

Method	Type	Bonn RGB-D Dynamic Dataset [47]					TUM RGB-D Dataset [62]				
		Ball	Ball2	Pers	Pers2	Avg.	f3/ws	f3/wx	f3/wr	f3/whs	Avg.
RGB-D Input											
ORB-SLAM2 [44]	S	6.5	23.0	6.9	7.9	11.1	40.8	72.2	80.5	72.3	66.45
NICE-SLAM [90]	S	24.4	20.2	24.5	53.6	30.7	79.8	86.5	244.0	152.0	140.57
ReFusion [48]	R	17.5	25.4	28.9	46.3	29.5	1.7	9.9	40.6	10.4	15.7
DynaSLAM (N+G) [4]	R	3.0	2.9	6.1	7.8	5.0	0.6	1.5	3.5	2.5	2.03
DG-SLAM [76]	R	3.7	4.1	4.5	6.9	4.8	0.6	1.6	4.3	-	-
RoDyn-SLAM [19]	R	7.9	11.5	14.5	13.8	11.9	1.7	8.3	-	5.6	-
DDN-SLAM (RGB-D) [33]	R	**1.8**	4.1	4.3	3.8	3.5	1.0	1.4	3.9	2.3	2.15
RGB Input											
DSO [10]	S	7.3	21.8	30.6	26.5	21.6	1.5	12.9	13.8	40.7	17.23
DROID-SLAM [66]	S	7.5	4.1	4.3	5.4	5.3	1.2	1.6	4.0	2.2	2.25
MonoGS [42]	S	15.3	17.3	26.4	35.2	23.6	1.1	21.5	17.4	44.2	21.05
Splat-SLAM [54]	S	8.8	3.0	4.9	25.8	10.6	2.3	1.3	3.9	2.2	2.43
DDN-SLAM (RGB) [33]	R	-	-	-	-	-	2.5	2.8	8.9	4.1	4.58
MegaSaM [37]	R	3.7	2.6	4.1	4.0	3.6	0.6	1.5	**2.6**	1.8	**1.63**
WildGS-SLAM [89]	R	2.7	**2.4**	3.6	3.1	**2.94**	**0.4**	1.3	3.3	**1.6**	**1.63**
DynaMoN (MS) [55]	D	6.8	3.8	**2.4**	3.5	4.1	1.4	1.4	3.9	2.0	2.18
DynaMoN (MS&SS) [55]	D	2.8	2.7	14.8	**2.2**	5.6	0.7	1.4	3.9	1.9	1.98
D4DGS-SLAM* [65]	D	3.6	3.9	4.5	5.2	4.3	-	-	-	-	-
4D-GS SLAM* [35]	D	2.4	3.7	8.9	9.4	6.1	0.5	2.1	**2.6**	-	-
ProDyG (Ours)	D	2.7	2.6	4.9	2.9	3.29	1.6	**1.2**	3.0	1.7	1.89

Tracking: Quantitative Results (ATE RMSE [cm])

- Type S: Static scenes, Type R: Robust against dynamics, Type D: Dynamic reconstruction

Method	Type	Bonn RGB-D Dynamic Dataset [47]					TUM RGB-D Dataset [62]				
		Ball	Ball2	Pers	Pers2	Avg.	f3/ws	f3/wx	f3/wr	f3/whs	Avg.
RGB-D Input											
ORB-SLAM2 [44]	S	6.5	23.0	6.9	7.9	11.1	40.8	72.2	80.5	72.3	66.45
NICE-SLAM [90]	S	24.4	20.2	24.5	53.6	30.7	79.8	86.5	244.0	152.0	140.57
ReFusion [48]	R	17.5	25.4	28.9	46.3	29.5	1.7	9.9	40.6	10.4	15.7
DynaSLAM (N+G) [4]	R	3.0	2.9	6.1	7.8	5.0	0.6	1.5	3.5	2.5	2.03
DG-SLAM [76]	R	3.7	4.1	4.5	6.9	4.8	0.6	1.6	4.3	-	-
RoDyn-SLAM [19]	R	7.9	11.5	14.5	13.8	11.9	1.7	8.3	-	5.6	-
DDN-SLAM (RGB-D) [33]	R	**1.8**	4.1	4.3	3.8	3.5	1.0	1.4	3.9	2.3	2.15
RGB Input											
DSO [10]	S	7.3	21.8	30.6	26.5	21.6	1.5	12.9	13.8	40.7	17.23
DROID-SLAM [66]	S	7.5	4.1	4.3	5.4	5.3	1.2	1.6	4.0	2.2	2.25
MonoGS [42]	S	15.3	17.3	26.4	35.2	23.6	1.1	21.5	17.4	44.2	21.05
Splat-SLAM [54]	S	8.8	3.0	4.9	25.8	10.6	2.3	1.3	3.9	2.2	2.43
DDN-SLAM (RGB) [33]	R	-	-	-	-	-	2.5	2.8	8.9	4.1	4.58
MegaSaM [37]	R	3.7	2.6	4.1	4.0	3.6	0.6	1.5	**2.6**	1.8	**1.63**
WildGS-SLAM [89]	R	2.7	**2.4**	3.6	3.1	**2.94**	**0.4**	1.3	3.3	**1.6**	**1.63**
DynaMoN (MS) [55]	D	6.8	3.8	**2.4**	3.5	4.1	1.4	1.4	3.9	2.0	2.18
DynaMoN (MS&SS) [55]	D	2.8	2.7	14.8	**2.2**	5.6	0.7	1.4	3.9	1.9	1.98
D4DGS-SLAM* [65]	D	3.6	3.9	4.5	5.2	4.3	-	-	-	-	-
4D-GS SLAM* [35]	D	2.4	3.7	8.9	9.4	6.1	0.5	2.1	**2.6**	-	-
ProDyG (Ours)	D	2.7	2.6	4.9	2.9	3.29	1.6	**1.2**	3.0	1.7	1.89

Tracking: Quantitative Results (ATE RMSE [cm])

- Type S: Static scenes, Type R: Robust against dynamics, Type D: Dynamic reconstruction
- *ProDyG* best among type **D**, and only slightly worse than WildGS-SLAM (**R**)

Method	Type	Bonn RGB-D Dynamic Dataset [47]					TUM RGB-D Dataset [62]				
		Ball	Ball2	Pers	Pers2	Avg.	f3/ws	f3/wx	f3/wr	f3/whs	Avg.
RGB-D Input											
ORB-SLAM2 [44]	S	6.5	23.0	6.9	7.9	11.1	40.8	72.2	80.5	72.3	66.45
NICE-SLAM [90]	S	24.4	20.2	24.5	53.6	30.7	79.8	86.5	244.0	152.0	140.57
ReFusion [48]	R	17.5	25.4	28.9	46.3	29.5	1.7	9.9	40.6	10.4	15.7
DynaSLAM (N+G) [4]	R	3.0	2.9	6.1	7.8	5.0	0.6	1.5	3.5	2.5	2.03
DG-SLAM [76]	R	3.7	4.1	4.5	6.9	4.8	0.6	1.6	4.3	-	-
RoDyn-SLAM [19]	R	7.9	11.5	14.5	13.8	11.9	1.7	8.3	-	5.6	-
DDN-SLAM (RGB-D) [33]	R	**1.8**	4.1	4.3	3.8	3.5	1.0	1.4	3.9	2.3	2.15
RGB Input											
DSO [10]	S	7.3	21.8	30.6	26.5	21.6	1.5	12.9	13.8	40.7	17.23
DROID-SLAM [66]	S	7.5	4.1	4.3	5.4	5.3	1.2	1.6	4.0	2.2	2.25
MonoGS [42]	S	15.3	17.3	26.4	35.2	23.6	1.1	21.5	17.4	44.2	21.05
Splat-SLAM [54]	S	8.8	3.0	4.9	25.8	10.6	2.3	1.3	3.9	2.2	2.43
DDN-SLAM (RGB) [33]	R	-	-	-	-	-	2.5	2.8	8.9	4.1	4.58
MegaSaM [37]	R	3.7	2.6	4.1	4.0	3.6	0.6	1.5	**2.6**	1.8	**1.63**
WildGS-SLAM [89]	R	2.7	**2.4**	3.6	3.1	**2.94**	**0.4**	1.3	3.3	**1.6**	**1.63**
DynaMoN (MS) [55]	D	6.8	3.8	**2.4**	3.5	4.1	1.4	1.4	3.9	2.0	2.18
DynaMoN (MS&SS) [55]	D	2.8	2.7	14.8	**2.2**	5.6	0.7	1.4	3.9	1.9	1.98
D4DGS-SLAM* [65]	D	3.6	3.9	4.5	5.2	4.3	-	-	-	-	-
4D-GS SLAM* [35]	D	2.4	3.7	8.9	9.4	6.1	0.5	2.1	**2.6**	-	-
ProDyG (Ours)	D	2.7	2.6	4.9	2.9	3.29	1.6	**1.2**	3.0	1.7	1.89

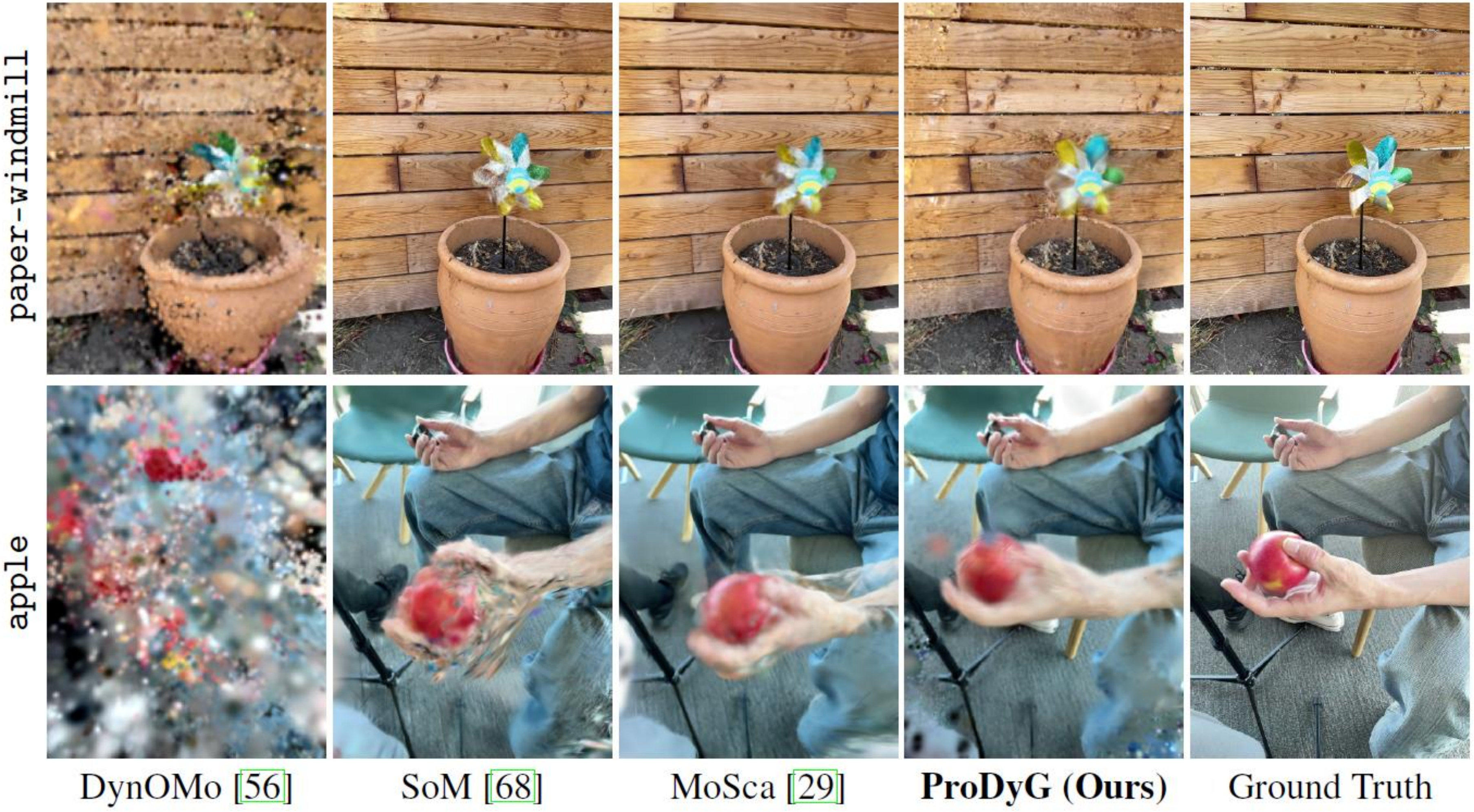
Novel View Synthesis: Quantitative Results

- Evaluated on the iPhone dataset [Gao 2022] from fixed test views
- *ProDyG* tested under 4 settings: with / without online tracking, RGB-D / RGB-only
- Beats Shape of Motion [Wang 2024] in PSNR and SSIM, and only marginally worse than MoSca [Lei 2024], **with extra constraints of online reconstruction and tracking**
- Significantly better than DynOMo [Seidenschwarz 2024] which is the only online baseline method
- RGB-only still works reasonably well

	Shape of Motion [68]	DynOMo [56]	MoSca [29]	Gaussian Marbles [61]	**ProDyG (Ours)**	**ProDyG (Ours)**	**ProDyG (Ours)**	**ProDyG (Ours)**
Online Reconstr.	✗	✓	✗	✗	✓	✓	✓	✓
Online Tracking	✗	✗	✗	✗	✗	✓	✗	✓
RGB-only	✗	✗	✗	✗	✗	✗	✓	✓
PSNR↑	17.43	11.98	**18.44**	16.00	17.65	17.87	15.41	15.40
SSIM ↑	0.591	0.436	**0.666**	-	0.634	0.643	0.603	0.582
LPIPS↓	**0.303**	0.748	0.311	0.437	0.390	0.377	0.462	0.492

Novel View Synthesis: Qualitative Results

- Evaluated on the iPhone dataset [Gao 2022]
- Dynamic objects reconstructed by *ProDyG* tend to show more accurate silhouettes

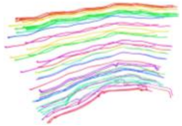


More Visual Results as Videos (Bonn RGB-D)

RGB Input

Depth Input

Est. Camera Trajectory



Dyn. Gaussian
Trajectories

Training View Rerendering

Fixed Novel View Rendering + Estimated Camera Poses + Dynamic Gaussian Trajectories

ETH zürich

Thank you for listening!

Project page: https://cs-vision.github.io/ProDyG.github.io

or scan: