LEARNING TO SEGMENT DYNAMIC OBJECTS USING SLAM OUTLIERS

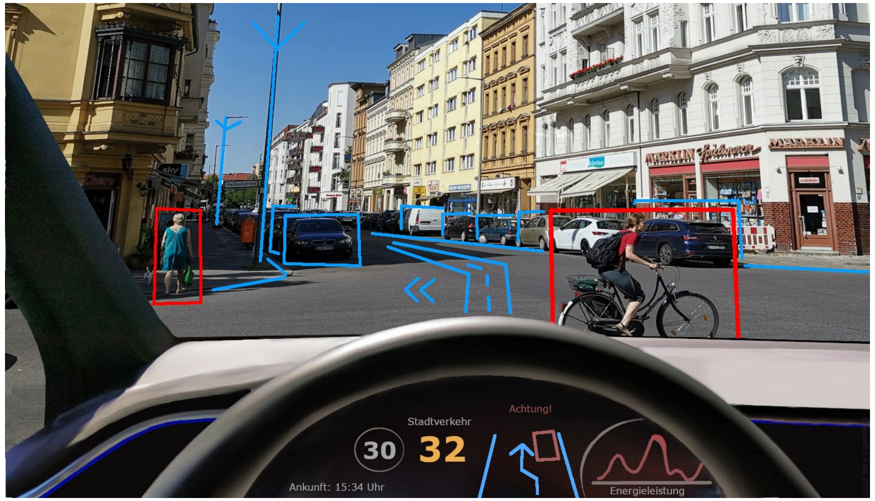
Adrian Bojko, Romain Dupont, Mohamed Tamaazousti and Hervé Le Borgne
Paris-Saclay University, CEA, List, France

adrian.bojko@cea.fr - romain.dupont@cea.fr
mohamed.tamaazousti@cea.fr - herve.le-borgne@cea.fr

ICPR 2020 PRESENTATION

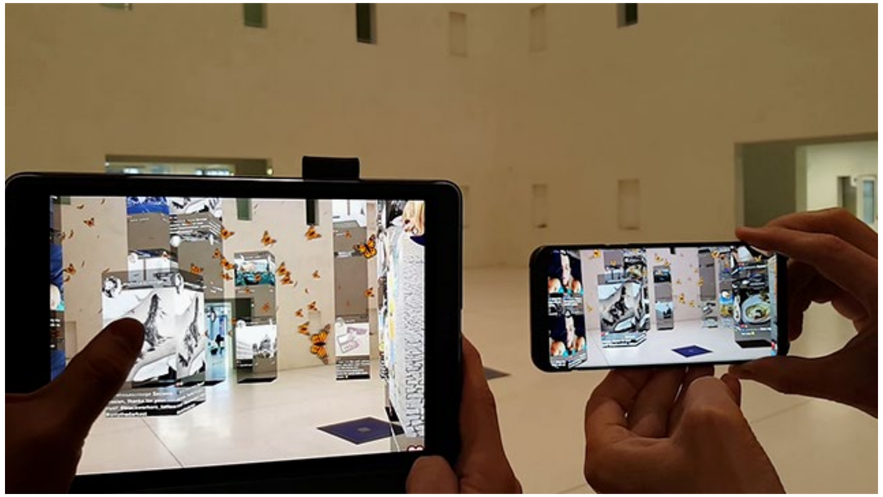
1. CONTEXT

1. **SLAM: Simultaneous Localization and Mapping in a static environment**
2. **Dynamic SLAM: SLAM extended to dynamic environments**
3. **Some applications:**



Autonomous vehicles

Robotics

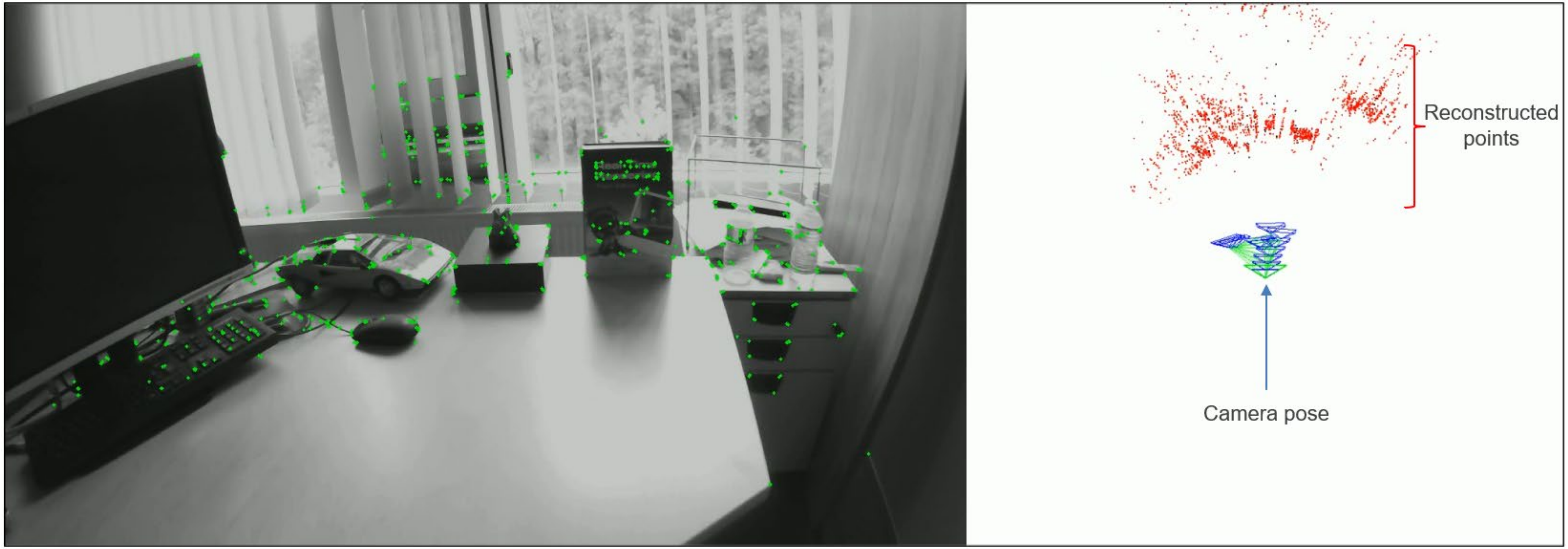


Augmented reality

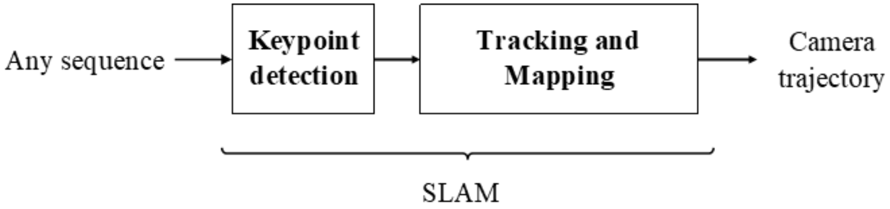
Images adapted from Wikimedia.

1. FOCUS: FEATURE-BASED SLAM

We use ORB-SLAM 2 [Mur-Artal et al., 2016].



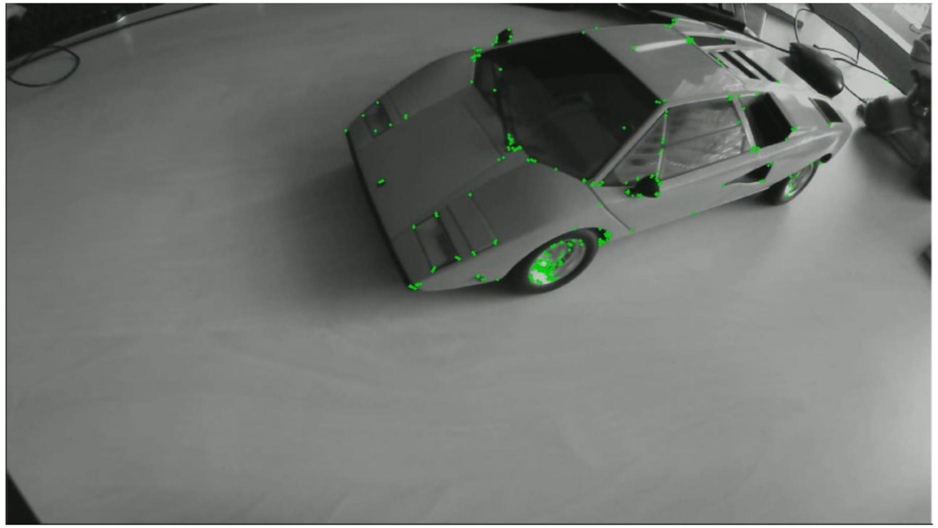
Sequence + keypoints (in green)   **Trajectory + map**



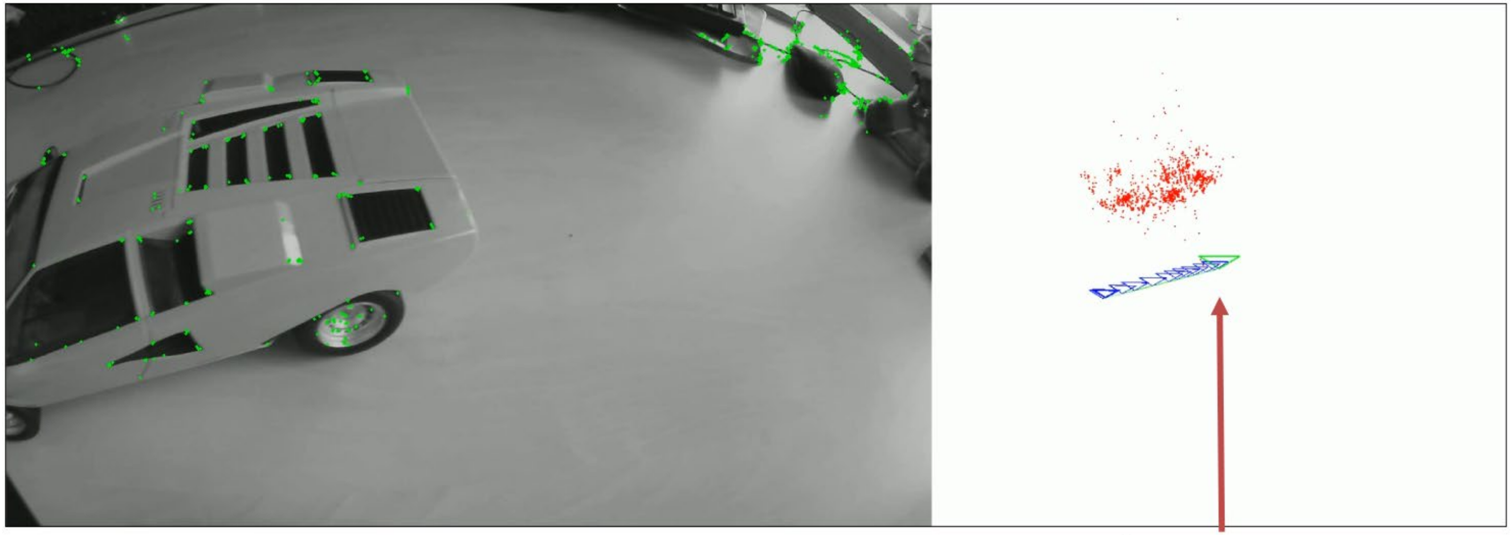
Feature-based SLAM: generate and track keypoints (in green) across images to compute camera trajectory while reconstructing the environment.

1. PROBLEM: CONSENSUS INVERSION

Consensus Inversion: implicit use of a frame of reference that is not the ground when the motion of dynamic objects is dominant.



Motion start (keypoints in green)



Motion end and final map

Camera pose

Example of false start (a type of consensus inversion)**:**
the camera is static but ORB-SLAM 2 (monocular) computes a fake trajectory.

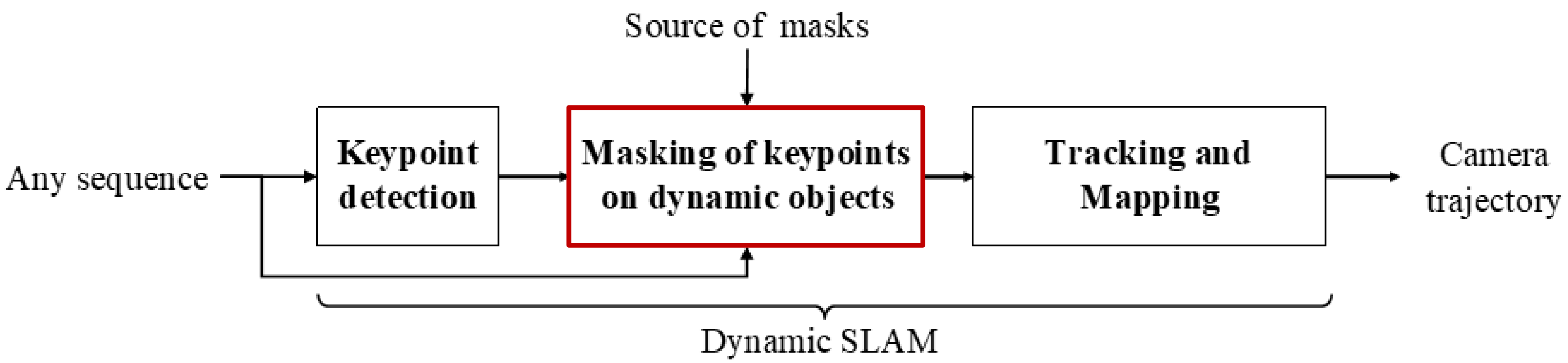
1. PROBLEM: PRIORS ON DYNAMIC OBJECTS CAN BE WRONG

Priors on dynamic objects (e.g. people) can be completely wrong.
The train station is moving, not the people nor the train.
Scene from "Top Secret !" (1984, Paramount Pictures)

2. STATE OF THE ART: DYNAMIC SLAMS

General principle of Dynamic SLAM: filter interest points on dynamic objects

→ **Critical step:** dynamic object detection



2. STATE OF THE ART: DYNAMIC SLAMS

Approaches	Limits
1. SLAM + geometry: → Uses: optical flow, depth maps... *An Accurate Localization Scheme* [Chen et al., 2018]	**→ Vulnerable to consensus inversions**
2. SLAM + semantic masks: → Uses: Mask R-CNN, ... *Mask-SLAM* [Kaneko et al., 2018]	**→ Limited by the scope of the training databases**
3. Hybrid: **a) SLAM + geometry (runtime) + semantic masks (runtime)** *DynaSLAM* [Bescos et al., 2018], *SLAMANTIC* [Schorghuber et al., 2019]	**→ Vulnerable to consensus inversions**
b) SLAM + geometry (training) + semantic masks (runtime) *Driven to distraction* [Barnes et al., 2018]	**→ Requires a lot of training data** (several traversals of the same location)

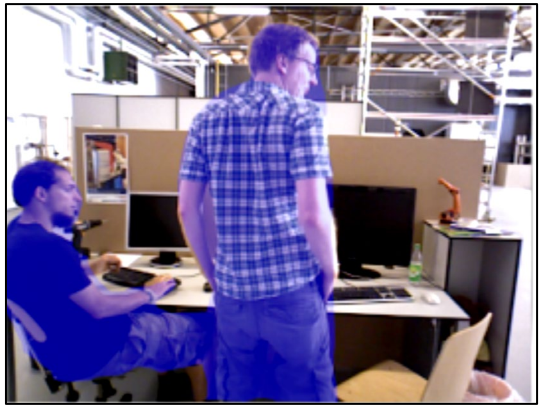
3. CONTRIBUTIONS

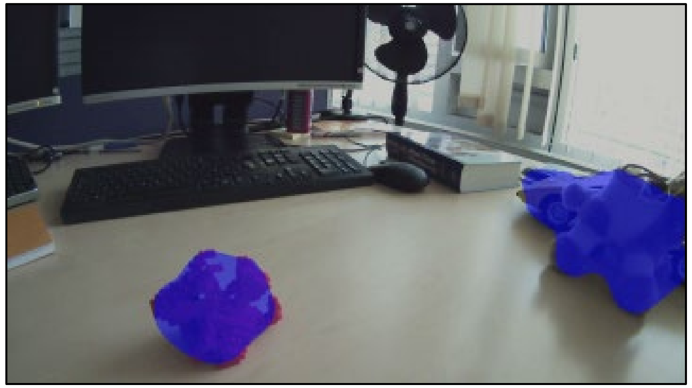
Our main contribution is a Dynamic SLAM:

- Based on self-supervised learning of masks
  (we use outliers i.e. keypoints rejected during optimization)
- Supports consensus inversions
- That only requires one learning sequence per dynamic object

Additional contributions:
1) Database *Consensus Inversion*

2) SLAM Robustness metrics (*Penalized ATE RMSE* and *Success Rate*)

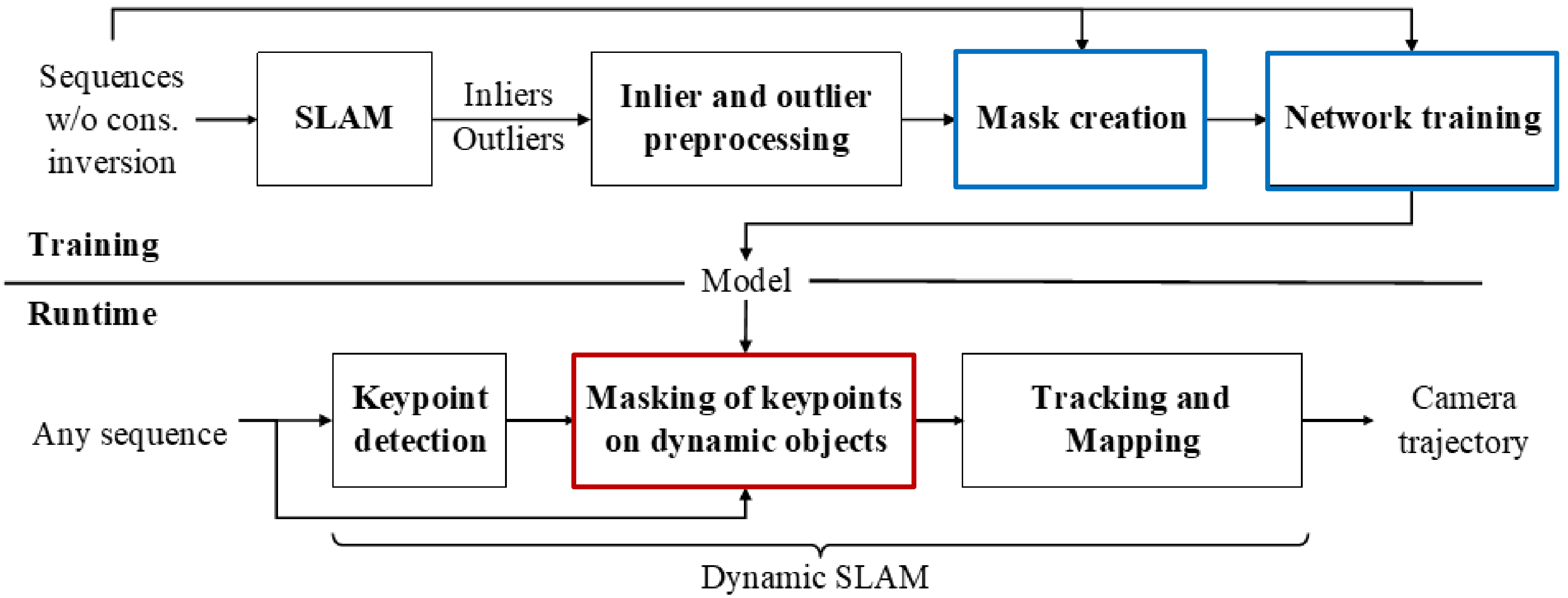






Images from TUM RGB-D and Consensus Inversion dataset, masked with our method.

3.1 METHOD

Hypothesis: dense outliers that appear suddenly characterize dynamic objects in sequences with no consensus inversion.



Dynamic SLAM = SLAM + semantic filter of keypoints

3.1 METHOD: MASK CREATION

- **SLAM inlier / outlier collection**
- **Mask database creation:**

  a) Search for dense outliers using sliding windows + creation of bounding boxes
     - We look for drops in the inlier/outlier ratio inside the sliding window.
     - We then merge overlapping boxes that have inlier/outlier ratio drops. The result is bounding boxes encloding dynamic boxes.

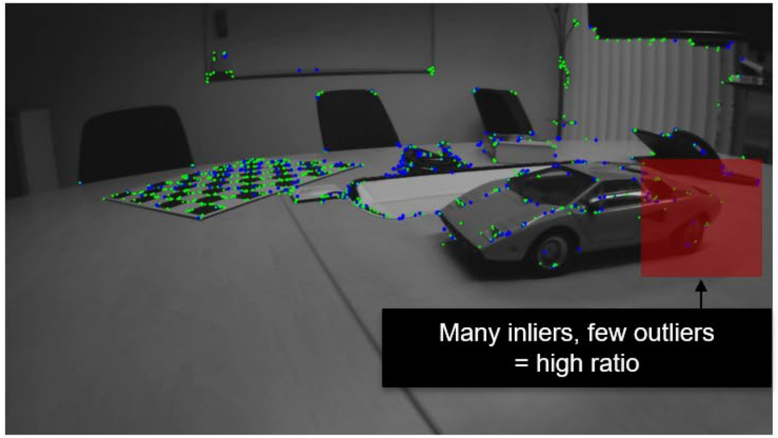


Image n: <u>before</u> the car moves.
(inliers in green, outliers in blue)

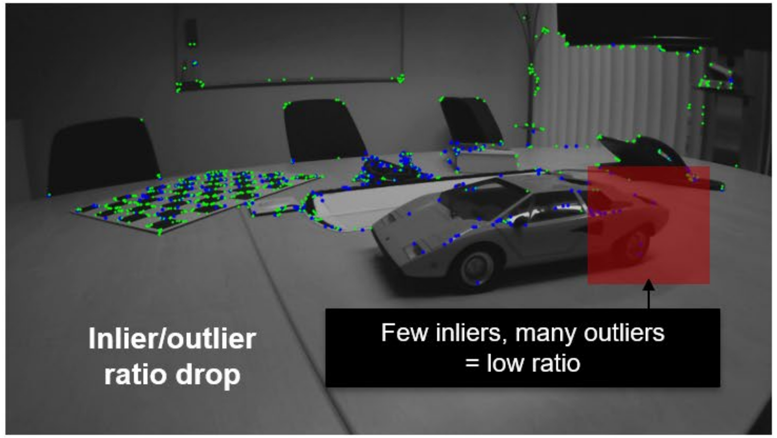


Image $n+3$: <u>after</u> the car moves.

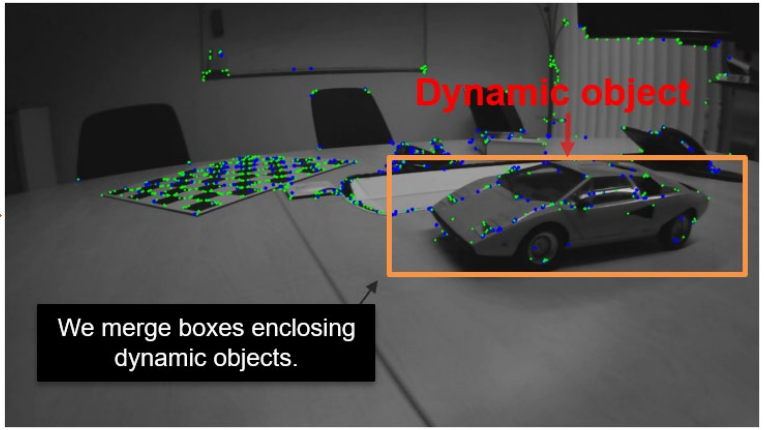


Image n: after merging windows with inlier/outlier ratio drops.

  b) Creation and propagation of masks across sequences using video segmentation tools:
     COSNet [Lu et al, 2019] and *SiamMask* [Ventura et al., 2019]

Dynamic object segmented in the whole sequence

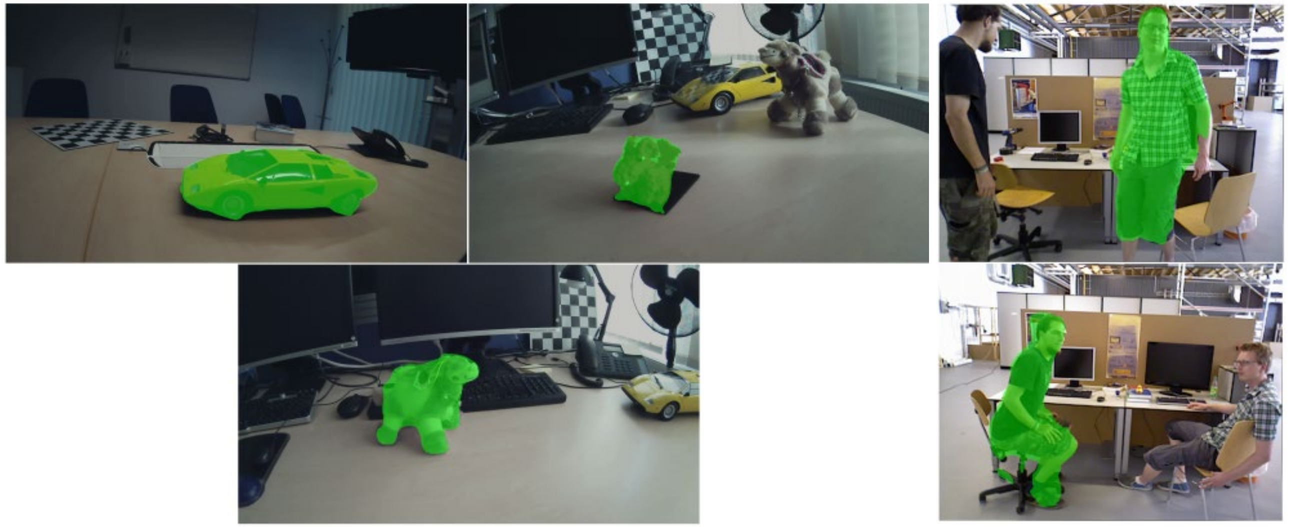
3.2 METHOD: NETWORK TRAINING

a) Train single-object models using the created mask database, DeepLabv3+ [Chen et al., 2018] architecture

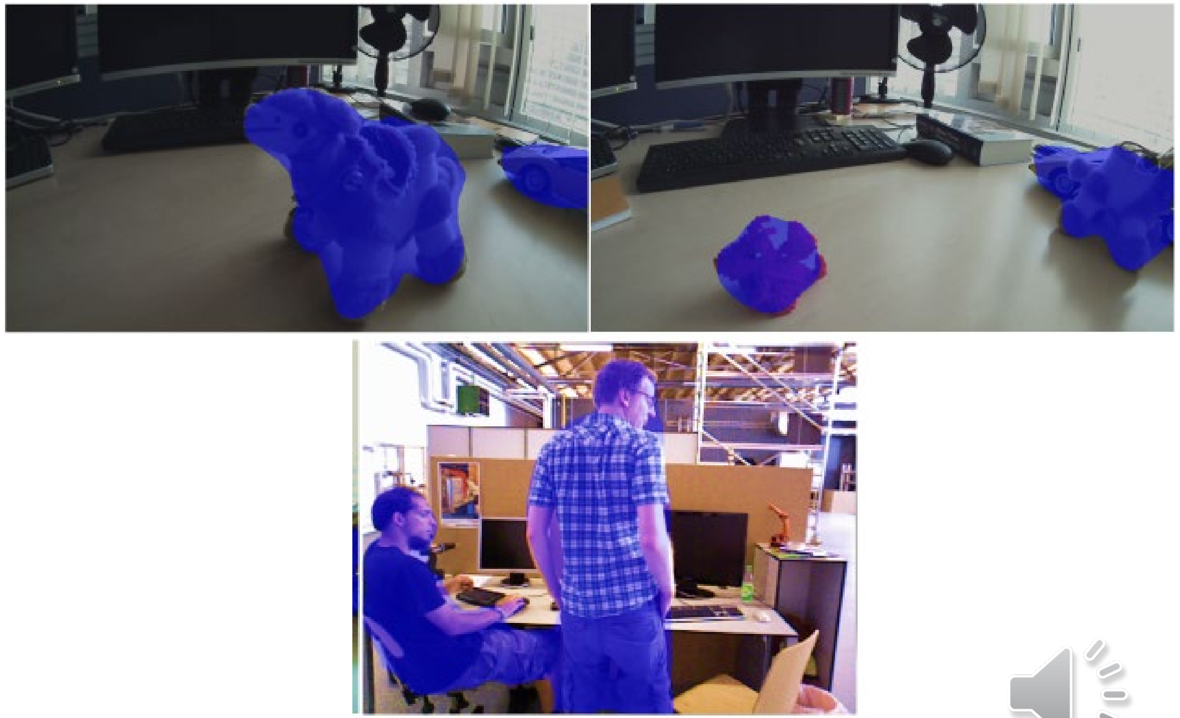
b) Infer masks with each model and superimpose the result per sequence

c) Train a global model with the superimposed masks

Single-object models: mask objects separately



Global model: mask all objects simultaneously



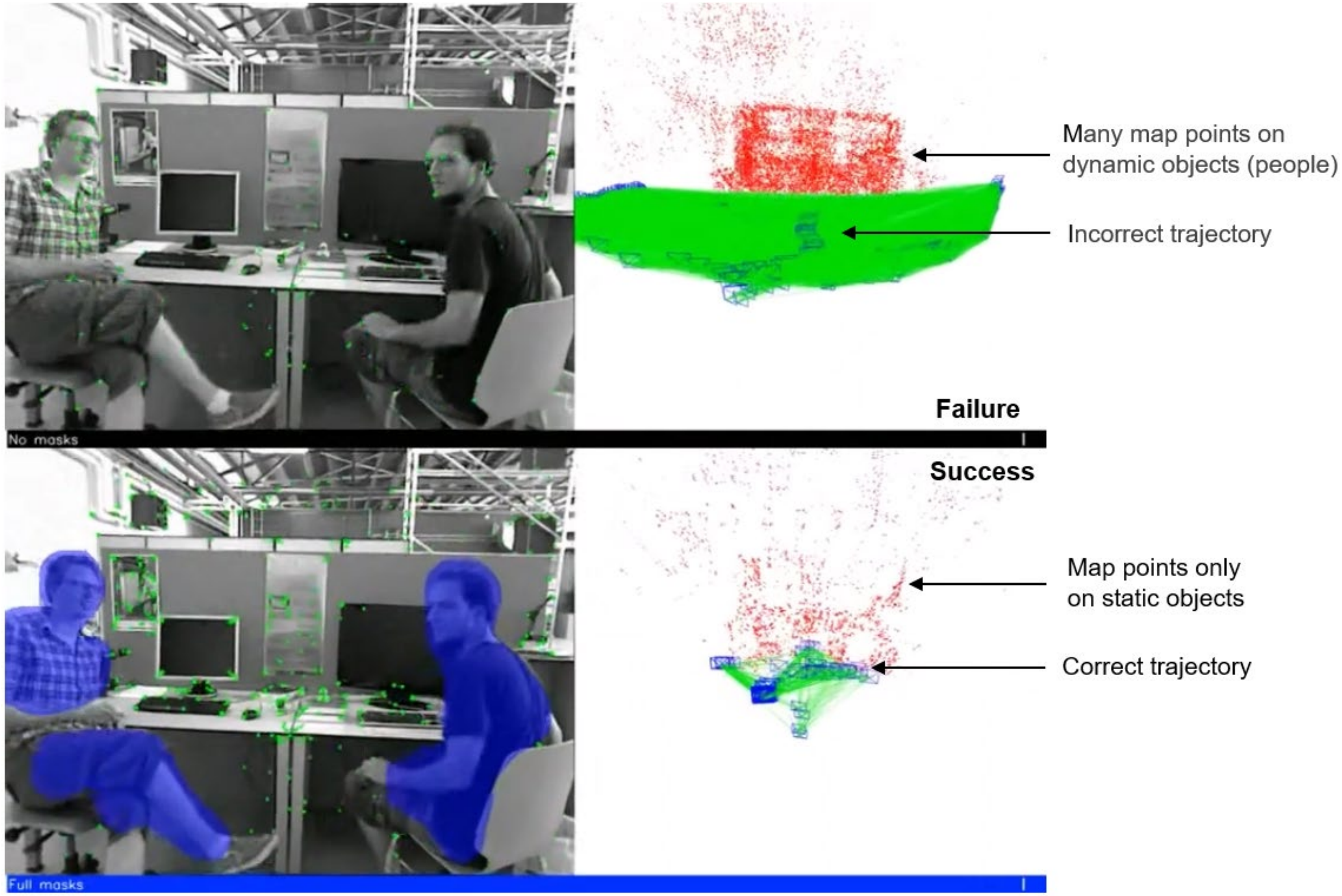
3.3 ADDITIONAL CONTRIBUTIONS

- **Our dataset "Consensus Inversion"** contains sequences with consensus inversion, rarely present in SLAM datasets.

- **Metrics to measure SLAM robustness:**
  - **SLAM failure:** Tracking Rate too low (compared to perfectly masking moving objects and consensus inversions) or ATE RMSE above a fixed threshold (e.g. 10cm).
  - $$\textbf{Penalized ATE RMSE} = \begin{cases} \max(L).(1+\tau), & \text{if SLAM failure} \\ \text{ATE RMSE} & \text{otherwise} \end{cases}$$
    - Defined within a SLAM benchmark.
    - L is the set of ATE RMSEs of all benchmarked SLAMs that were successful and τ the penalty factor.
  - **Success Rate:** % of sequences that are successfully processed by the SLAM.

3.4 QUALITATIVE RESULTS

- Example on TUM RGB-D [Sturm et al., 2012] (a popular SLAM database) in RGB-D
- Moving people cause a consensus inversion



Sequence *fr3_walking_xyz* (RGB-D), tracked keypoints in green.

3.4 QUANTITATIVE RESULTS

- **Evaluation on TUM RGB-D** (dynamic sequences) and **Consensus Inversion**
- Also tested network integration in LDSO [Gao et al, 2018], a direct SLAM
- Results (partial):

Test set	State-of-the-Art		ORB-SLAM 2 + ...				
			Segmentation baselines				
	DynaSLAM	SLAMANTIC	No seg.	Mask R-CNN	RVOS [Ventura et al. 2019]	COSNet	Our seg. [Bojko et al., 2020]
Consensus Inversion / Dyn. - Mono	0.0693	0.0692	0.0860	0.0760	0.0144	0.0297	**0.0089**
TUM RGB-D / Dyn. - Mono	0.1108	0.1101	0.0252	0.0235	0.0331	0.0267	**0.0222**
Consensus Inversion / Dyn. - Stereo	0.0627	0.0699	0.0756	0.0630	0.0116	0.0148	**0.0094**
TUM RGB-D / Dyn. - RGB-D	0.0206	0.0173	0.1077	**0.0172**	0.0218	0.0245	0.0185

Average Penalized ATE RMSE (m)

Better or equal accuracy than the state of the art.

Test set	State-of-the-Art		ORB-SLAM 2 + ...				
			Segmentation baselines				
	DynaSLAM	SLAMANTIC	No seg,	Mask R-CNN	RVOS	COSNet	Our seg.
Consensus Inversion / Dyn. - Mono	63,6%	63,6%	45,5%	54,5%	72,7%	72,7%	**100.0%**
TUM RGB-D / Dyn. - Mono	62,5%	62,5%	87,5%	87,5%	62,5%	**100.0%**	**100.0%**
Consensus Inversion / Dyn. - Stereo	72,7%	63,6%	63,6%	63,6%	81,8%	81,8%	**100.0%**
TUM RGB-D / Dyn. - RGB-D	**100.0%**	**100.0%**	62,5%	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Success Rate (%)

We prevent all SLAM failures in all modes.

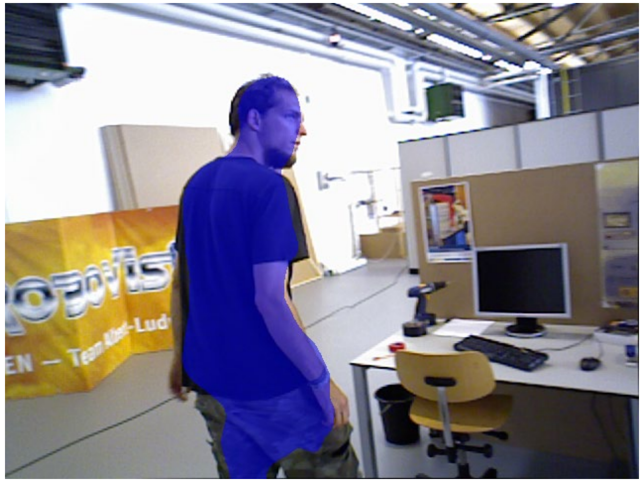
	LDSO + …	
	No seg,	Our seg.
Avg. Penalized ATE RMSE (m)	0.0833	**0.0581**
Success Rate (%)	36.4%	**63.6%**

LDSO on Consensus Inversion / Dyn.

Significant improvements on LDSO.

Results are better or equal than the state of the art in Mono / Stereo / RGB-D.

CONCLUSION



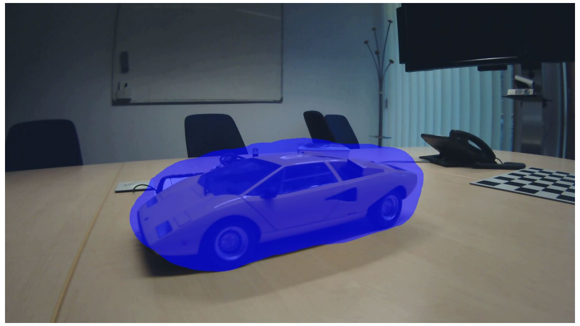
Contributions:

1. A novel method that learns to segment dynamic objects
   - No manual labelling.
   - Uses only one monocular sequence per dynamic object.
   - Supports consensus inversions.

2. The first dataset for Consensus Inversion evaluation.
3. The first robustness metrics that integrate SLAM failures.

Results:

- We improved ORB-SLAM 2 monocular/stereo/RGB-D as well as LDSO and achieved top results in very challenging scenarios.



BIBLIOGRAPHY

"*ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras*" (Mur-Artal et al., IEEE Transactions on Robotics 2017)

"*An Accurate Localization Scheme for Mobile Robots Using Optical Flow in Dynamic Environments*" (Cheng et al., ROBIO 2018)

"*Mask-SLAM: Robust Feature-Based Monocular SLAM by Masking Using Semantic Segmentation*" (Kaneko et al., CVPRW 2018)

"*DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes*" (Bescos et al., IEEE Robotics and Automation Letters, 2018)

"*Slamantic - leveraging semantics to improve vslam in dynamic environments*" (Schorghuber et al. ICCV Worshops, 2019).

"*Driven to Distraction: Self-Supervised Distractor Learning for Robust Monocular Visual Odometry in Urban Environments*" (Barnes et al., ICRA 2018)

"*See More, Know More: Unsupervised Video Object Segmentation With Co-Attention Siamese Networks*" (Lu et al., CVPR 2019)

"*Fast Online Object Tracking and Segmentation: A Unifying Approach*" (Wang et al., CVPR 2019)

"*Rvos: End-to-end recurrent network for video object segmentation*" (Ventura et al., CVPR 2019)

"*Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation*" (Chen et al., ECCV 2018)

"LDSO: Direct Sparse Odometry with Loop Closure" (Gao et al., IROS 2018)