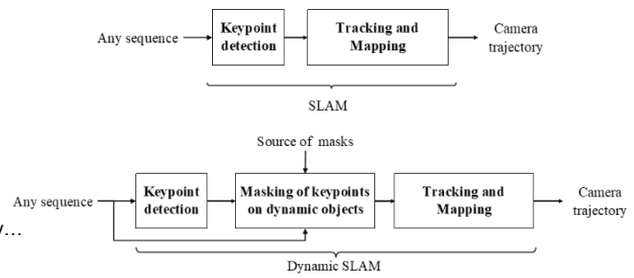
LEARNING TO SEGMENT DYNAMIC OBJECTS USING SLAM OUTLIERS

Adrian BOJKO, Romain DUPONT, Mohamed TAMAAZOUSTI, Hervé LE BORGNE

Paris-Saclay University, CEA, List, F-91120, Palaiseau, France

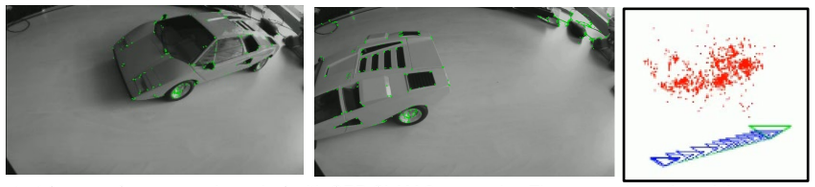
SLAM and Dynamic SLAM

- **SLAM** = Simultaneous Localization and Mapping
  - Feature-based: track keypoints across images.
- **Dynamic SLAM** = SLAM in Dynamic environments
  - Filter keypoints on detected dynamic objects.
  - The detection step is crucial.
- **Used in** Robotics, Autonomous Vehicles, Augmented Reality…



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.
  - May happen if an object moves when the camera is very close to it.
  - Very difficult to detect with geometric methods since they rely on the dominant motion of the image.
- **The SLAM may significantly drift or even compute fake trajectories due to false starts.**



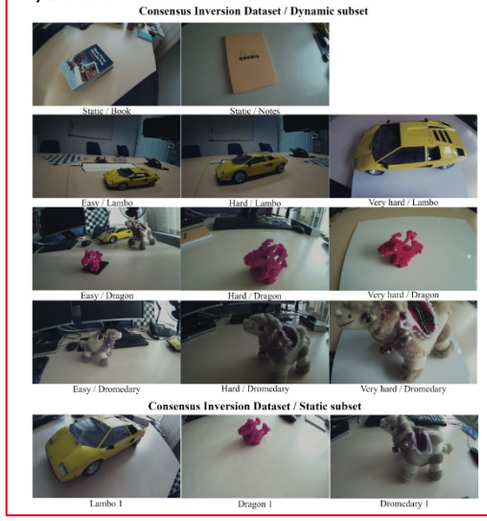
Example of false start (a case of consensus inversion) with ORB-SLAM 2 monocular. The camera is static and the car moves from right to left.
Left: before car motion. **Middle**: after car motion. **Right**: final SLAM map (red) and computed poses (blue).
The computed trajectory (right) is nonsense as there is no camera motion.

Contributions

Our main contribution is a Dynamic SLAM:

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

1) Database *Consensus Inversion*

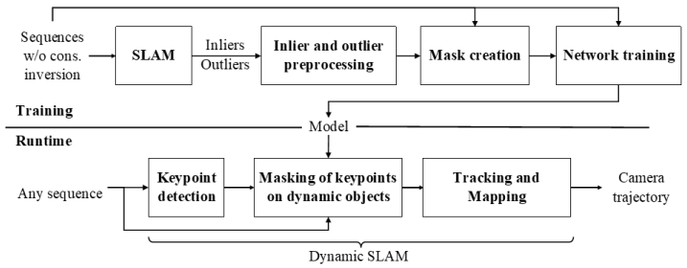


2) SLAM Robustness metrics

- Standard metrics are **ATE RMSE** (Absolute Trajectory Error) and **Tracking Rate** (% tracked images). These metrics must be analyzed together which makes comparisons difficult.
- We propose the **Penalized ATE RMSE** and the **Success Rate**. The metrics are relative within a SLAM benchmark.
- **SLAM Failure**: the Tracking Rate of the sequence is too low (compared to an ideal masking) or the ATE RMSE is above a threshold.
- **Penalized ATE RMSE** $= \begin{cases} \max(L) \cdot (1+\tau), & \text{if SLAM failure} \\ \text{ATE RMSE} & \text{otherwise} \end{cases}$
  - L is the set of ATE RMSEs of all evaluated SLAMs that successfully processed the tested sequence and τ the penalty factor.
- **Success Rate** = % of dataset sequences the SLAM successfully processed.

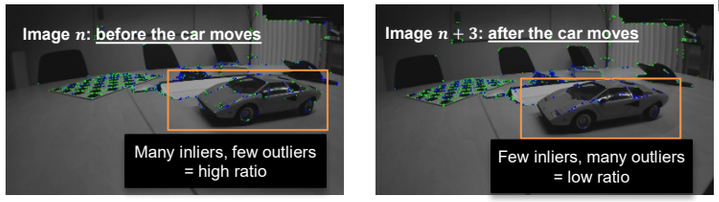
Method

- **Hypothesis:** dense outliers that appear suddenly characterize dynamic objects in sequences with no consensus inversion.
- **Dynamic SLAM** = SLAM + semantic filter of keypoints
- We use outliers and sequences without consensus inversion as input to the mask creation.



Mask Creation

a) Search for dense clusters of outliers using sliding windows + creation of bounding boxes



We search for drops in inlier/outlier ratio between images n and $n+3$ using sliding windows. When an object moves after being reconstructed by the SLAM, inliers (in green) on it are replaced with outliers (in blue) as long as there is no consensus inversion.

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.

Network training

a) We train a single-object model per sequence using the created mask database. DeepLabv3+ architecture [Chen et al., 2018].

b) We infer masks with each model and superimpose the result per sequence.

c) We use the superimposed masks to train a global model.

All dynamic objects are segmented simultaneously.

Results

- Our results are better or equal than the State of the Art on TUM RGB-D [Sturm et al., 2012] and Consensus Inversion in Monocular, Stereo and RGB-D.
- We prevent false starts and consensus inversions.

	State-of-the-Art		ORB-SLAM 2 + ...				
			Segmentation baselines				
Test set	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)

	State-of-the-Art		ORB-SLAM 2 + ...				
			Segmentation baselines				
Test set	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 (%)

	LDSO [Gao et al, 2018] + ...	
	No seg,	Our seg.
Avg. Penalized ATE RMSE (m)	0.0833	**0.0581**
Success Rate (%)	36.4%	**63.6%**

Conclusion

We proposed:

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

Additional contributions:

- The first dataset for Consensus Inversion evaluation.
- 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.