CVPR SEATTLE WASHINGTON
JUNE 16-18 2020

ClusterVO: Clustering Moving Instances and Estimating Visual Odometry for Self and Surroundings

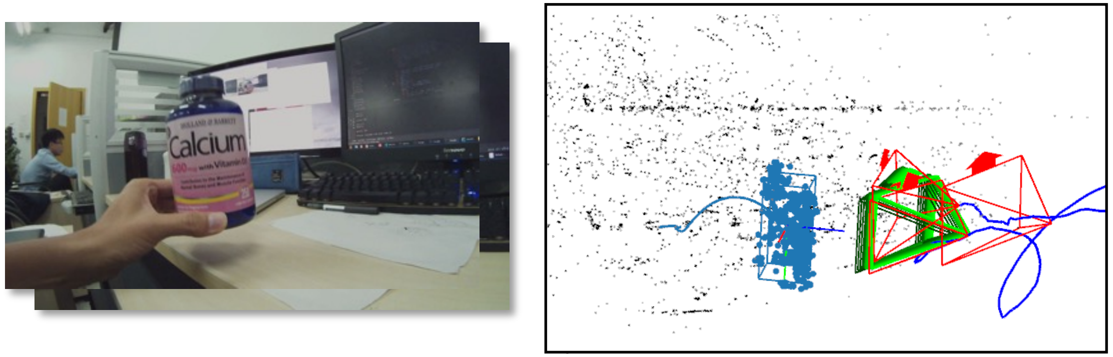
Jiahui Huang[1], Sheng Yang[2], Tai-Jiang Mu[1], Shi-Min Hu[1]

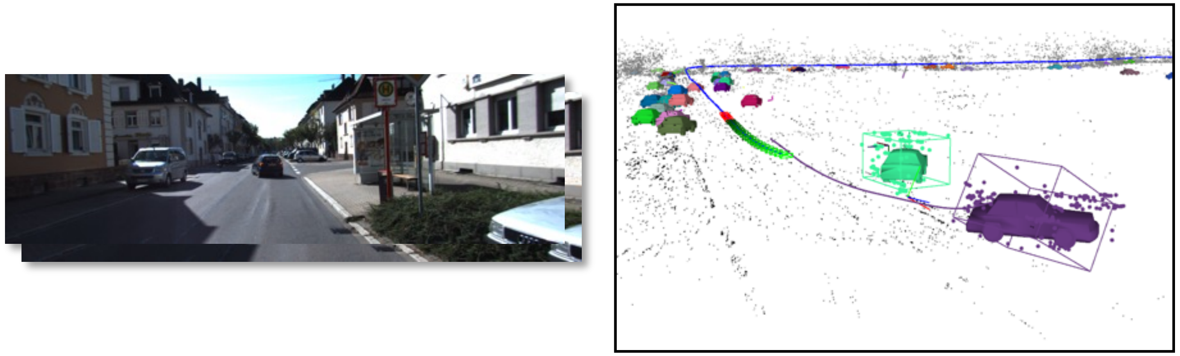
[1]BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing, China

[2]Alibaba Inc., China

Alibaba A.I. Labs
阿里巴巴人工智能实验室

Introduction

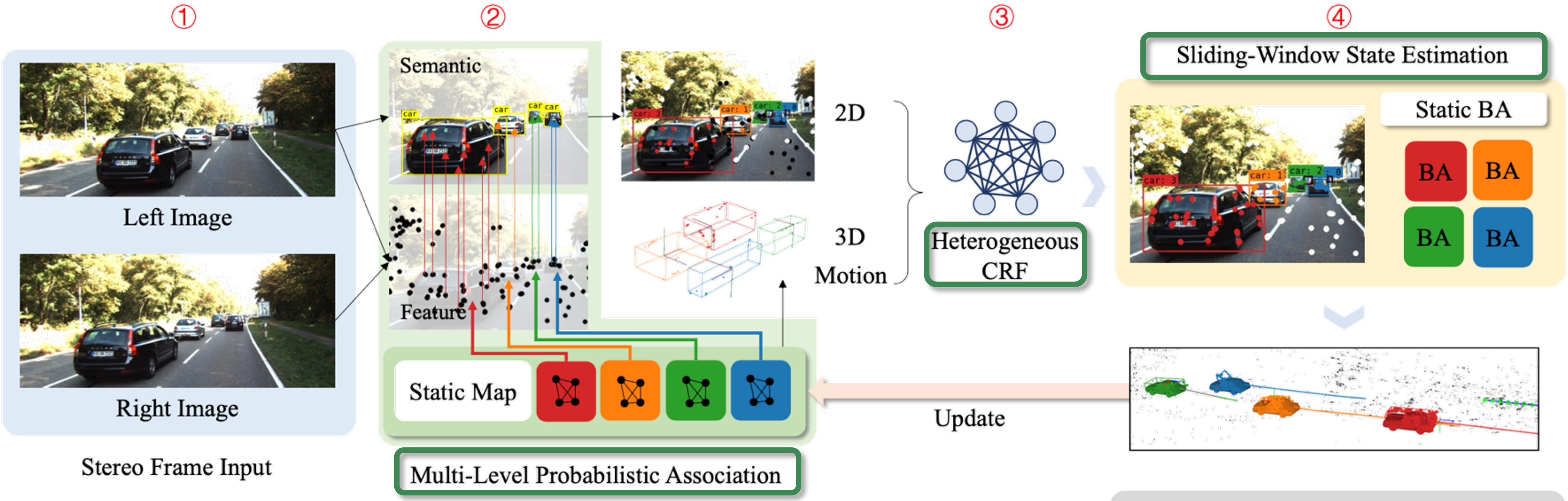




ClusterVO is a stereo Visual Odometry which simultaneously clusters and estimates the motion of both ego and surrounding rigid clusters/objects.

Unlike previous solutions relying on batch input or imposing priors on scene structure or dynamic object models, our method is online, general and applicable in various scenarios including indoor scene understanding and autonomous driving.

System Pipeline



①

Left Image

Right Image

Stereo Frame Input

②

Semantic

Feature

Static Map

Multi-Level Probabilistic Association

2D

3D Motion

③

Heterogeneous CRF

④

Sliding-Window State Estimation

Static BA

BA BA

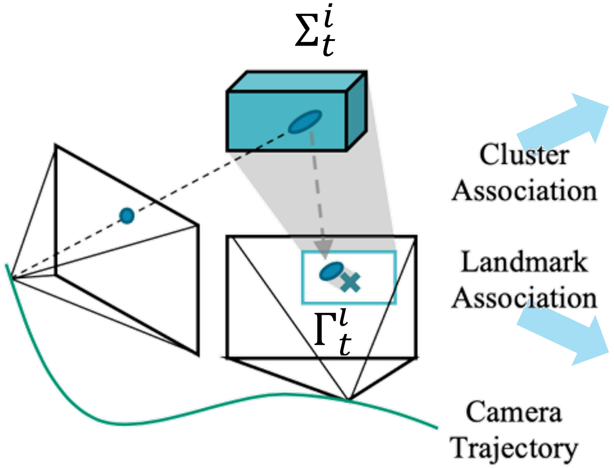
BA BA

Update

▭ Core Components

Multi-level Probabilistic Association

- For each new frame, we need to robustly associate detected features and semantic bounding boxes to map landmarks and clusters.
- Association probabilities for the two levels are calculated based on stereo triangulation uncertainty.



$$p_{\mathbf{q}}(m) \propto \sum_{\boldsymbol{\zeta}_t^k \in \boldsymbol{B}_t^m} (1/|\Gamma_t^i|)$$

$$p_i(k) \propto \left[\|\boldsymbol{\zeta}_t^i - \boldsymbol{z}_t^k\|_{\Gamma_t^i}^2 < \gamma \right] \cdot s_{ik}$$

$$\Sigma_t^i := \mathbf{R}_{t'}^{\mathbf{c}} \, {}_{\boldsymbol{Z}}\Sigma_{t'}^i \mathbf{R}_{t'}^{\mathbf{c}\top}, \quad t' := \operatorname*{argmin}_{t'<t} |\, {}_{\boldsymbol{Z}}\Sigma_{t'}^i|$$

$$\boldsymbol{\zeta}_t^i = \pi(\boldsymbol{p}_t^i + \boldsymbol{v}_t^{\mathbf{q}}) \qquad \Gamma_t^i = \mathbf{J}_\pi \Sigma_t^i \mathbf{J}_\pi^\top$$

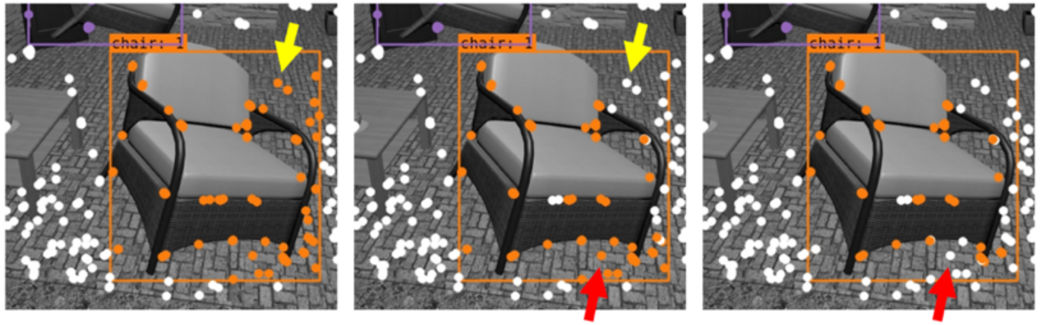
Heterogeneous CRF

- The cluster assignment $\mathbf{q}^i$ of each landmark i observed in the current frame is assigned by a Conditional Random Field model combining semantic (2D), spatial (3D) and motion information, which we call 'Heterogeneous CRF'.

$$E(\{\mathbf{q}^i\}_i) := \sum_i \psi_u(\mathbf{q}^i) + \alpha \sum_{i<j} \psi_p(\mathbf{q}^i, \mathbf{q}^j)$$

$$\psi_u(\mathbf{q}^i) \propto \boxed{p_{\mathrm{2D}}}(\mathbf{q}^i) \cdot \boxed{p_{\mathrm{3D}}}(\mathbf{q}^i) \cdot \boxed{p_{\mathrm{mot}}}(\mathbf{q}^i)$$

$$\psi_p(\mathbf{q}^i, \mathbf{q}^j) := [\mathbf{q}^i \neq \mathbf{q}^j] \cdot \exp(-\|\boldsymbol{p}_t^i - \boldsymbol{p}_t^j\|^2)$$



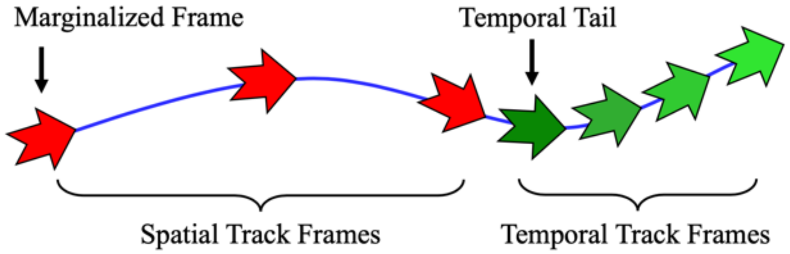
2D Only

2D + 3D

2D + 3D + Motion

Sliding-Window State Optimization

- We employ a novel double-track frame management design to maintain keyframes in dynamic scenes.
  - Temporal track frames allow for enough observations to track fast-moving clusters.
  - Spatial track frames help create enough parallax for accurate triangulation.



- The energy function for state optimization is defined separately for static and dynamic parts:

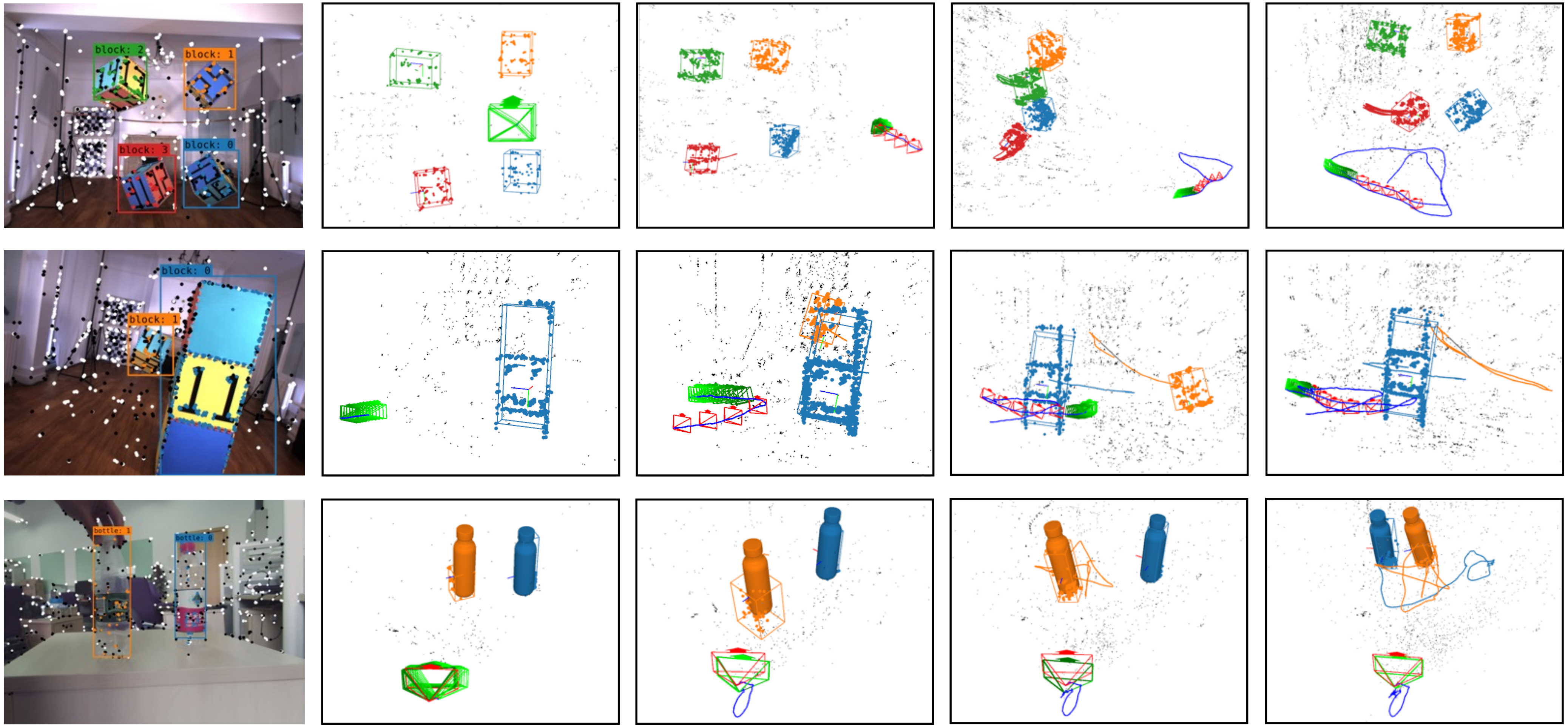
Static scene and Camera pose
BA Term + Marginalization

$$\mathbf{E}_s := \sum_{i\in\mathcal{I}_0, t\in\mathcal{T}_a} \rho(\|\boldsymbol{z}_t^i - \pi((\mathbf{P}_t^{\mathrm{c}})^{-1}\boldsymbol{p}_t^i)\|^2_{\boldsymbol{z}\Sigma}) + \sum_{t\in\mathcal{T}_a} \|\delta\boldsymbol{x}_t^{\mathbf{c}} - \mathbf{H}^{-1}\boldsymbol{\beta}\|^2_{\mathbf{H}},$$

Dynamic clusters
Dynamic BA Term +
Smoothness

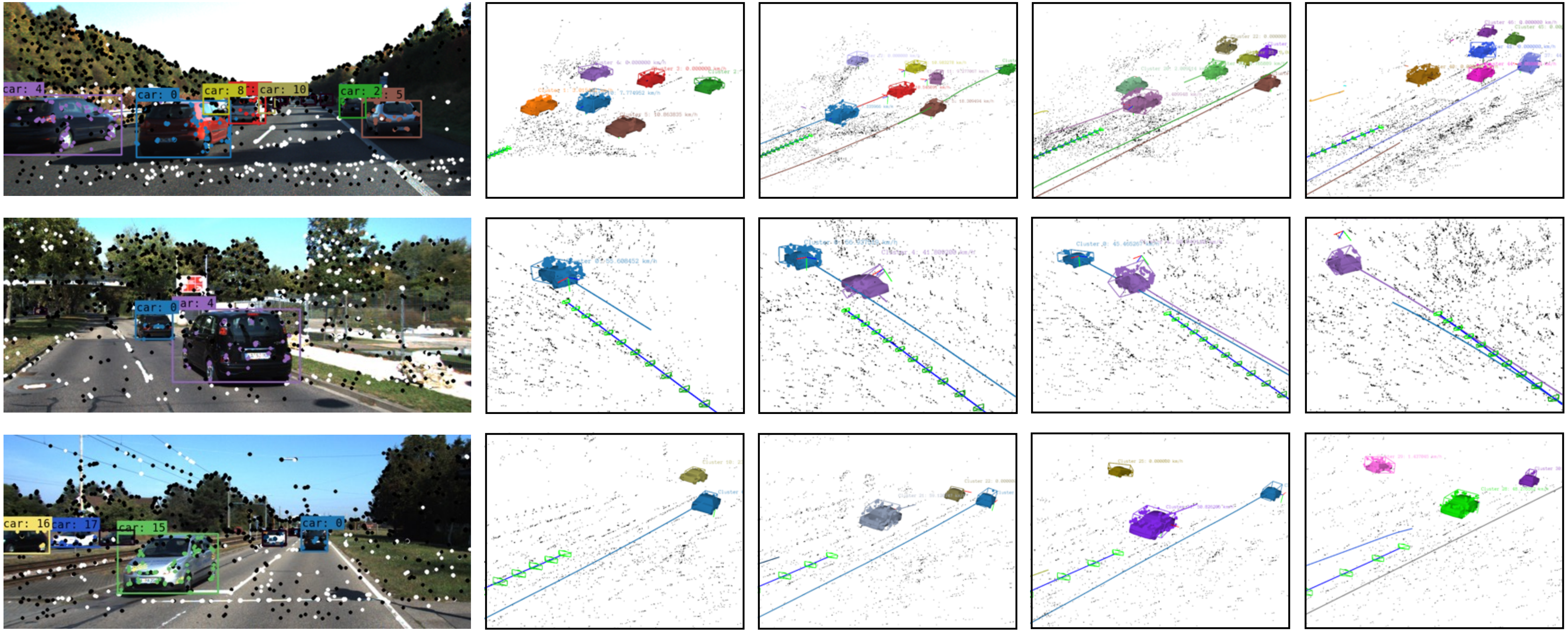
$$\mathbf{E}_d := \sum_{t,t^+\in\mathcal{T}_t} \left\| \begin{bmatrix} \boldsymbol{t}_{t^+}^i \\ \boldsymbol{v}_{t^+}^i \end{bmatrix} - \mathbf{A} \begin{bmatrix} \boldsymbol{t}_t^i \\ \boldsymbol{v}_t^i \end{bmatrix} \right\|^2_{\hat{Q}} + \sum_{i\in\mathcal{I}_{\mathbf{q}}, t\in\mathcal{T}_t} \rho(\|\boldsymbol{z}_t^i - \pi(\mathbf{T}_t^{\mathbf{cq}^i}(\mathbf{P}_t^{\mathbf{c}})^{-1}\boldsymbol{p}_t^i)\|^2_{\boldsymbol{z}\Sigma})$$

Results: Indoor Dataset



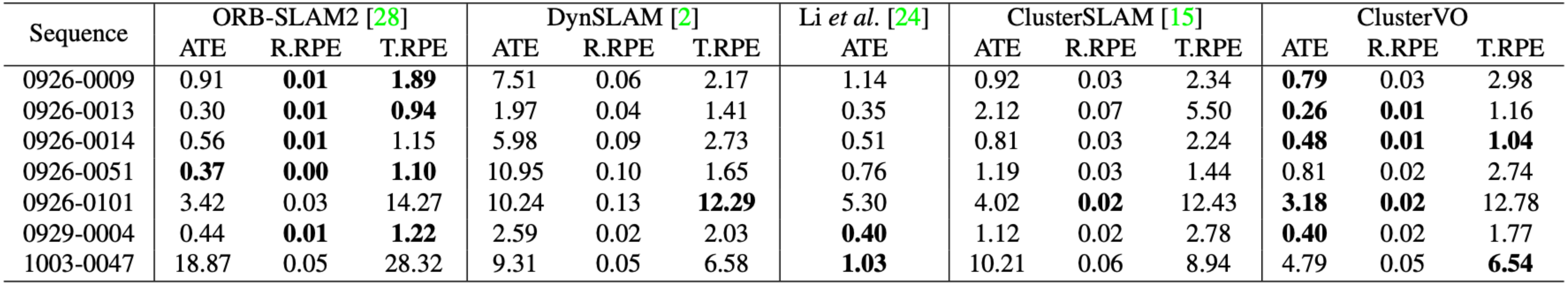
Results: KITTI Raw Dataset

CVPR SEATTLE WASHINGTON
JUNE 16-18 2020

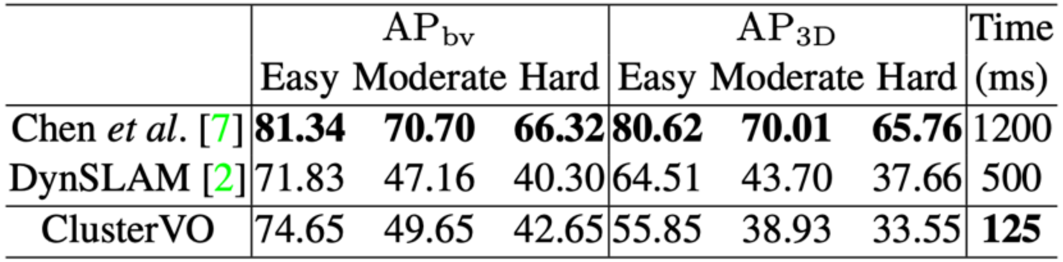


Comparisons

Ego-Motion on KITTI raw Dataset

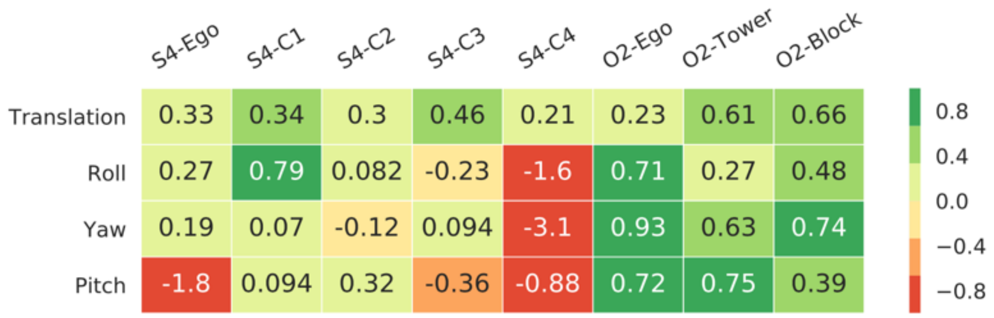


Sequence	ORB-SLAM2 [28]			DynSLAM [2]			Li *et al.* [24]	ClusterSLAM [15]			ClusterVO		
	ATE	R.RPE	T.RPE	ATE	R.RPE	T.RPE	ATE	ATE	R.RPE	T.RPE	ATE	R.RPE	T.RPE
0926-0009	0.91	**0.01**	**1.89**	7.51	0.06	2.17	1.14	0.92	0.03	2.34	**0.79**	0.03	2.98
0926-0013	0.30	**0.01**	**0.94**	1.97	0.04	1.41	0.35	2.12	0.07	5.50	**0.26**	**0.01**	1.16
0926-0014	0.56	**0.01**	1.15	5.98	0.09	2.73	0.51	0.81	0.03	2.24	**0.48**	**0.01**	**1.04**
0926-0051	**0.37**	**0.00**	**1.10**	10.95	0.10	1.65	0.76	1.19	0.03	1.44	0.81	0.02	2.74
0926-0101	3.42	0.03	14.27	10.24	0.13	**12.29**	5.30	4.02	**0.02**	12.43	**3.18**	**0.02**	12.78
0929-0004	0.44	**0.01**	**1.22**	2.59	0.02	2.03	**0.40**	1.12	0.02	2.78	**0.40**	0.02	1.77
1003-0047	18.87	0.05	28.32	9.31	0.05	6.58	**1.03**	10.21	0.06	8.94	4.79	0.05	**6.54**



	AP_{bv}			AP_{3D}			Time
	Easy	Moderate	Hard	Easy	Moderate	Hard	(ms)
Chen *et al.* [7]	**81.34**	**70.70**	**66.32**	**80.62**	**70.01**	**65.76**	1200
DynSLAM [2]	71.83	47.16	40.30	64.51	43.70	37.66	500
ClusterVO	74.65	49.65	42.65	55.85	38.93	33.55	**125**

Object Detection on KITTI



	S4-Ego	S4-C1	S4-C2	S4-C3	S4-C4	O2-Ego	O2-Tower	O2-Block
Translation	0.33	0.34	0.3	0.46	0.21	0.23	0.61	0.66
Roll	0.27	0.79	0.082	-0.23	-1.6	0.71	0.27	0.48
Yaw	0.19	0.07	-0.12	0.094	-3.1	0.93	0.63	0.74
Pitch	-1.8	0.094	0.32	-0.36	-0.88	0.72	0.75	0.39

Trajectory Accuracy on OMD

Thank You!