A Stereo Visual Odometry Framework with Augmented Perception for Dynamic Urban Environments

Marcelo Contreras[1], Neel. P Bhatt[2], and Ehsan Hashemi[2]

1 UTEC UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

2 UNIVERSITY OF ALBERTA

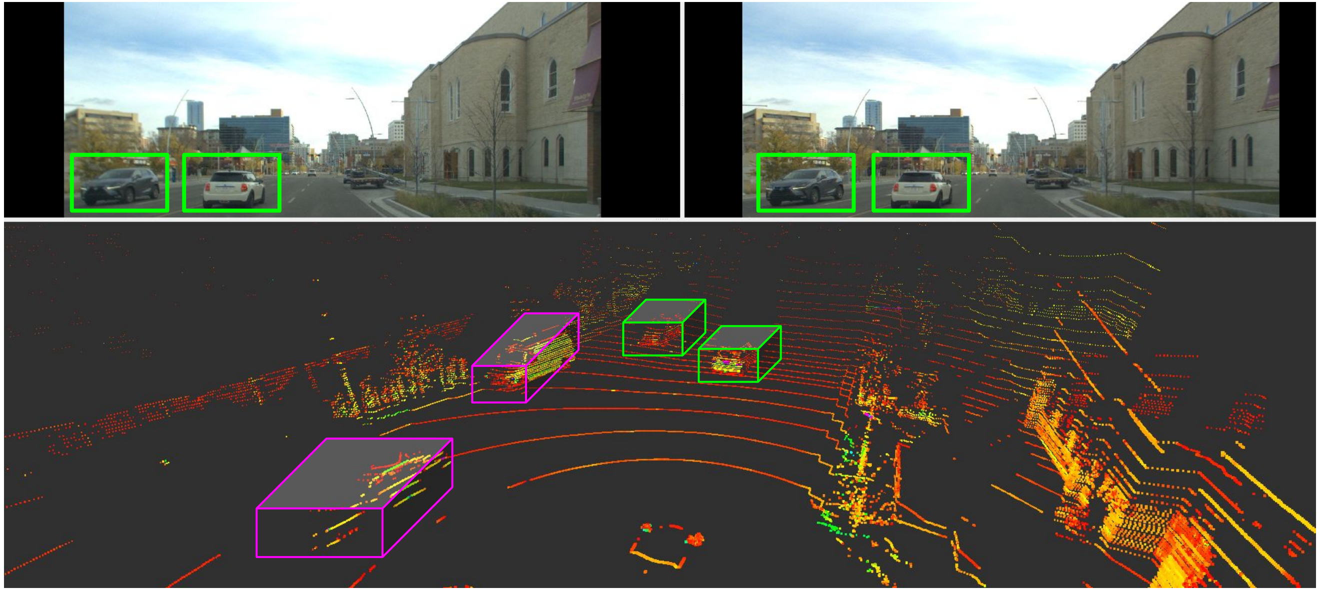
node LAB

IEEE ITSC 2023

Problem statement

node LAB

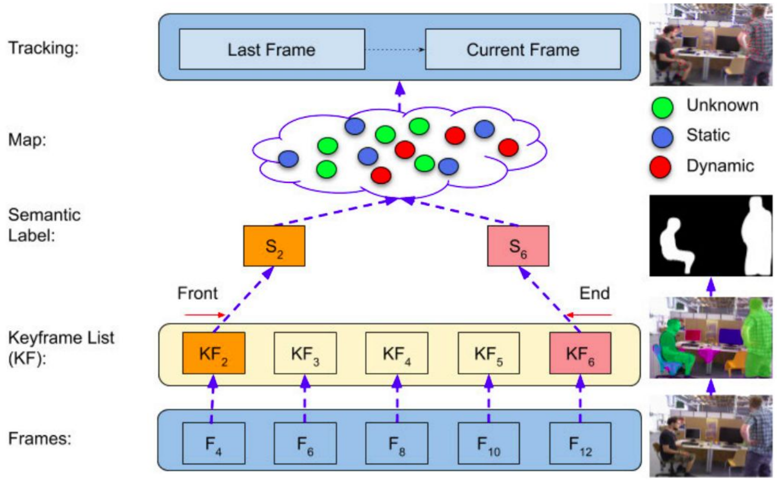
- In urban canyons, there is a significant presence of dynamic instances (vehicles and pedestrians) that generate untrackable landmarks for ego-motion estimation solutions.



Related works

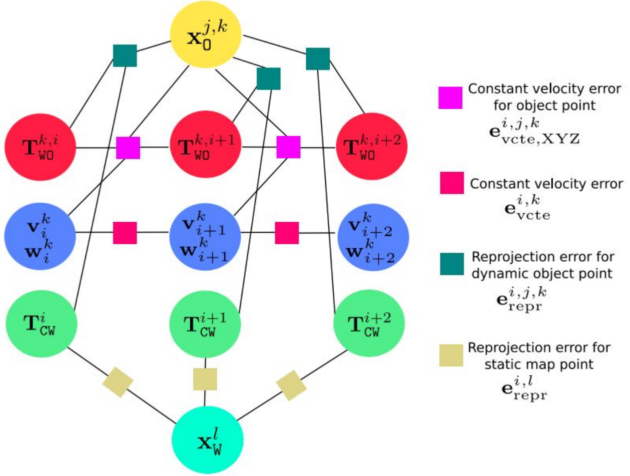
node LAB

Semantic based optimization



RDS-SLAM [1]

Tightly-Coupled Multi-Object Tracking and SLAM



DynaSLAM II [2]

[1] Y. Liu and J. Miura, "Rds-slam: Real-time dynamic slam using semantic segmentation methods," IEEE Access, vol. 9, pp. 23 772– 23 785, 2021.
[2] Bescos, B., Campos, C., Tardós, J. D., & Neira, J. (2021). DynaSLAM II: Tightly-Coupled Multi-Object Tracking and SLAM. IEEE robotics and automation letters, 6(3), 5191-5198. https://doi.org/10.1109/lra.2021.3068640

Our contributions

node LAB

We propose **dynamic stereo VO** system that integrates the following characteristics:

- Compute a **union-static mask** from a **priori static street instances** by merging instance segmentation and object detection.
    - Both features and landmarks are only associated to the **static background**.

- An **efficient bundle adjustment** over **semantic-aware feature tracking** for pose refinement over a moving horizon and a sparse set of static covisible landmarks.

Dynamic SVO system overview

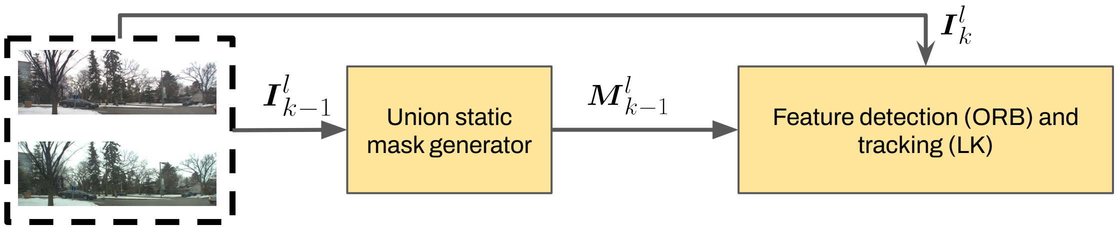
node LAB

We design the system that:

Dynamic SVO system overview

node LAB

1) Computes an union static mask from merged semantics of YOLACT and YOLOv5. The feature detection filters outliers outside the static mask, and inliers are tracked to current frame.

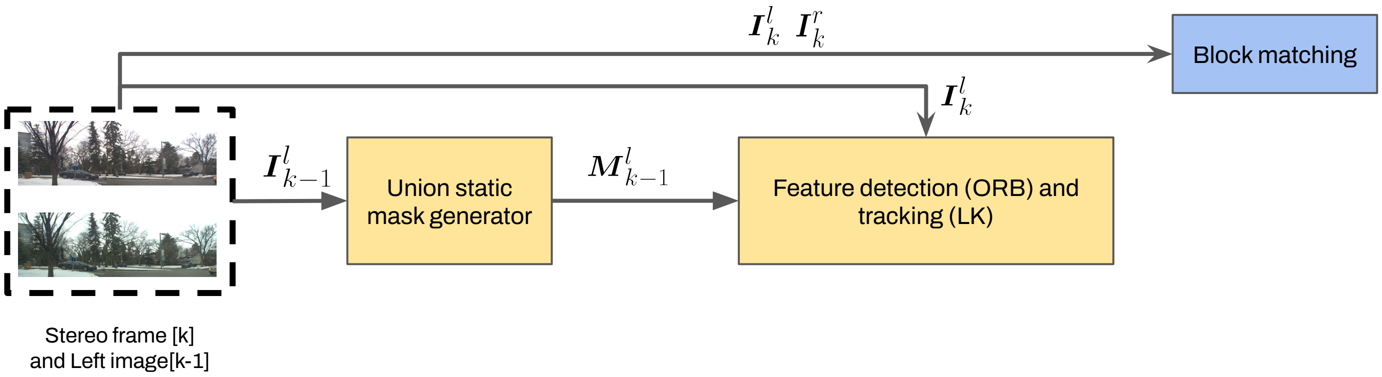


Stereo frame [k]
and Left image[k-1]

SVO algorithm overview

node LAB

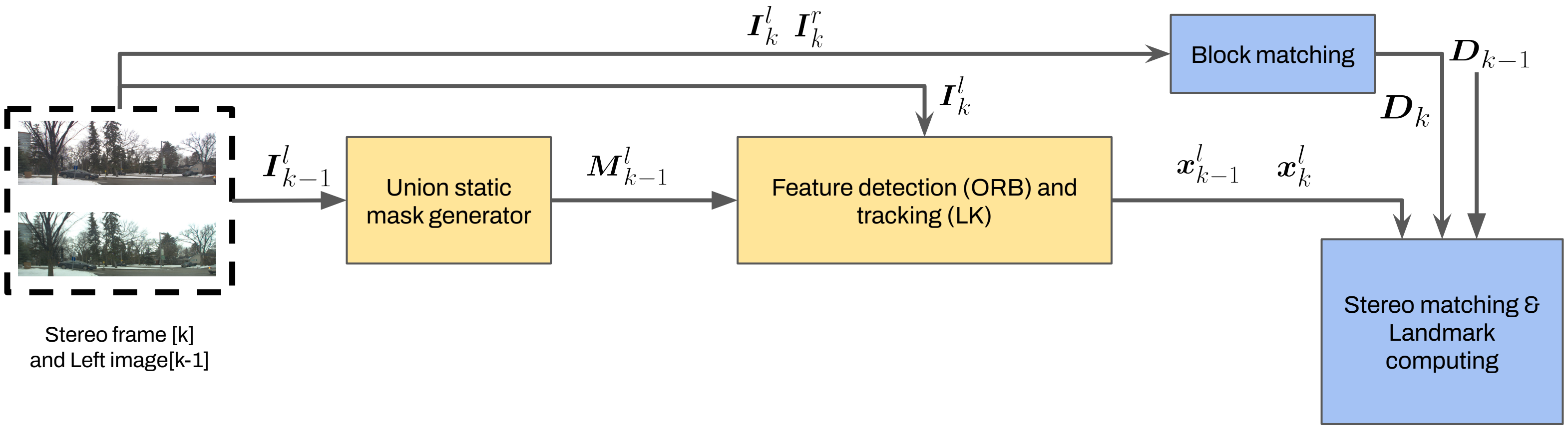
2) Block matches left and right frame to get disparity map and compute right features.



SVO algorithm overview

node LAB

2) Block matches left and right frame to get disparity map and right features without detection. Then, landmarks are triangulated to form a local point cloud.

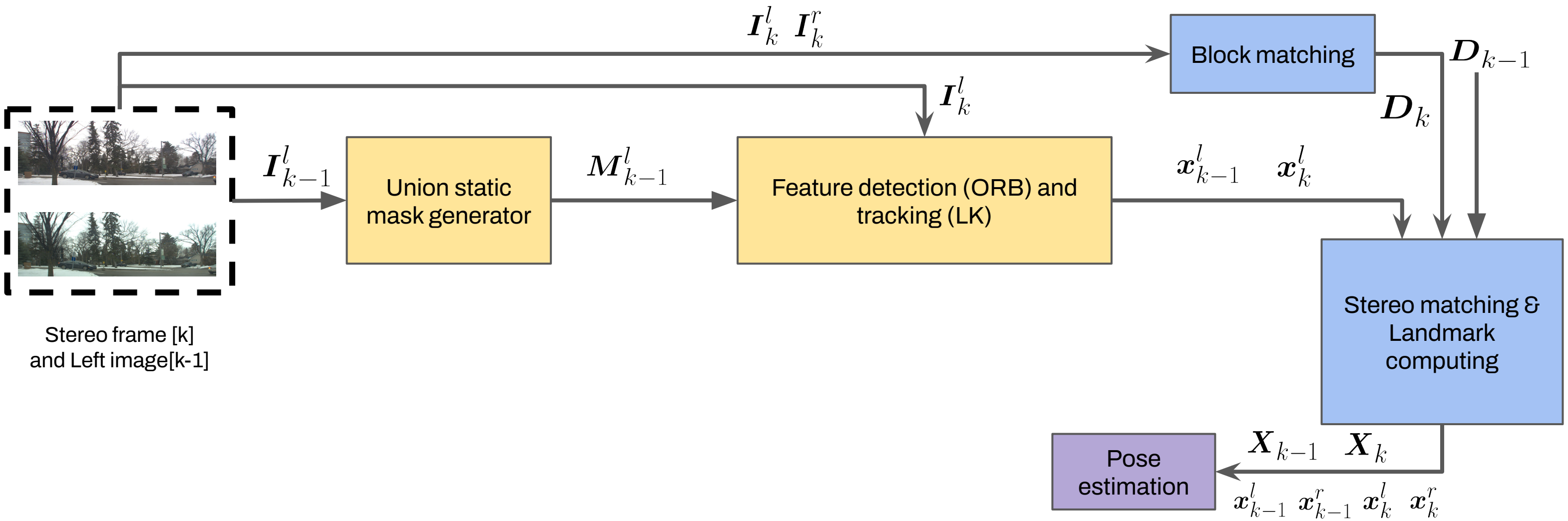


SVO algorithm overview

node LAB

2) Relative pose T (from k-1 to k) is estimated from optimization problem:

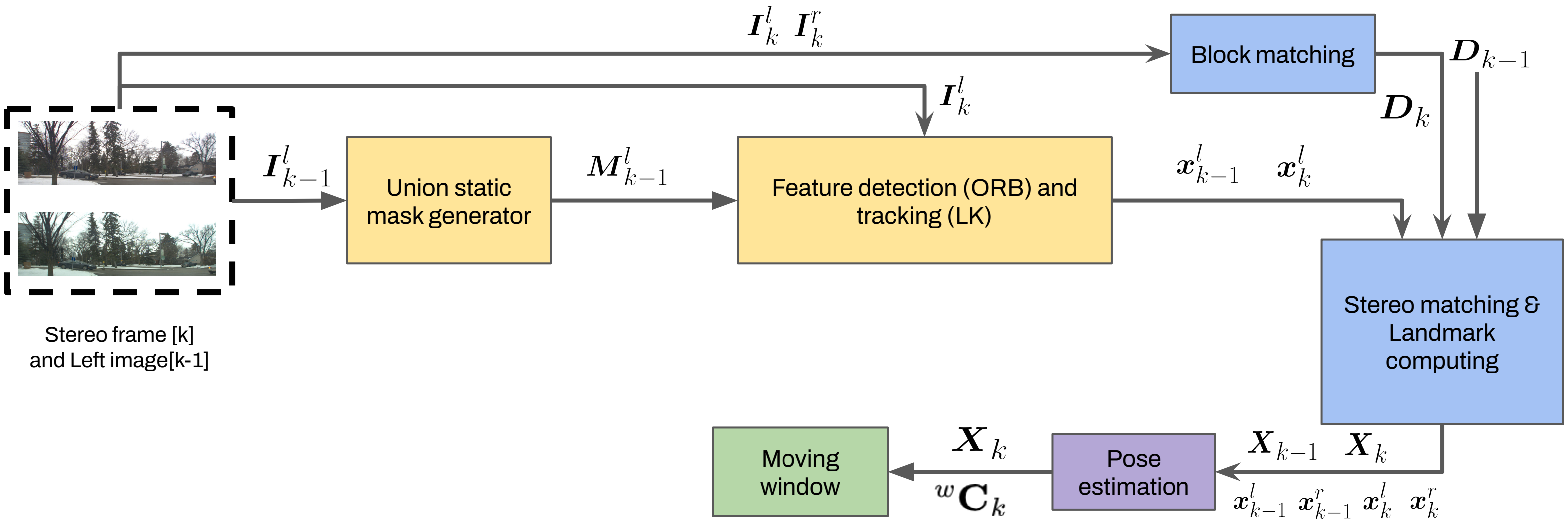
Eq(1)
$$\underset{{}^{k}\boldsymbol{T}_{k-1}}{\operatorname{argmin}} \left\| \boldsymbol{x}_k^l - \pi(\boldsymbol{X}_{k-1}, \boldsymbol{P}^l, {}^{k}\boldsymbol{T}_{k-1}) \right\|^2 + \left\| \boldsymbol{x}_{k-1}^l - \pi\left(\boldsymbol{X}_k, \boldsymbol{P}^l, {}^{k-1}\boldsymbol{T}_k\right) \right\|^2$$



SVO algorithm overview

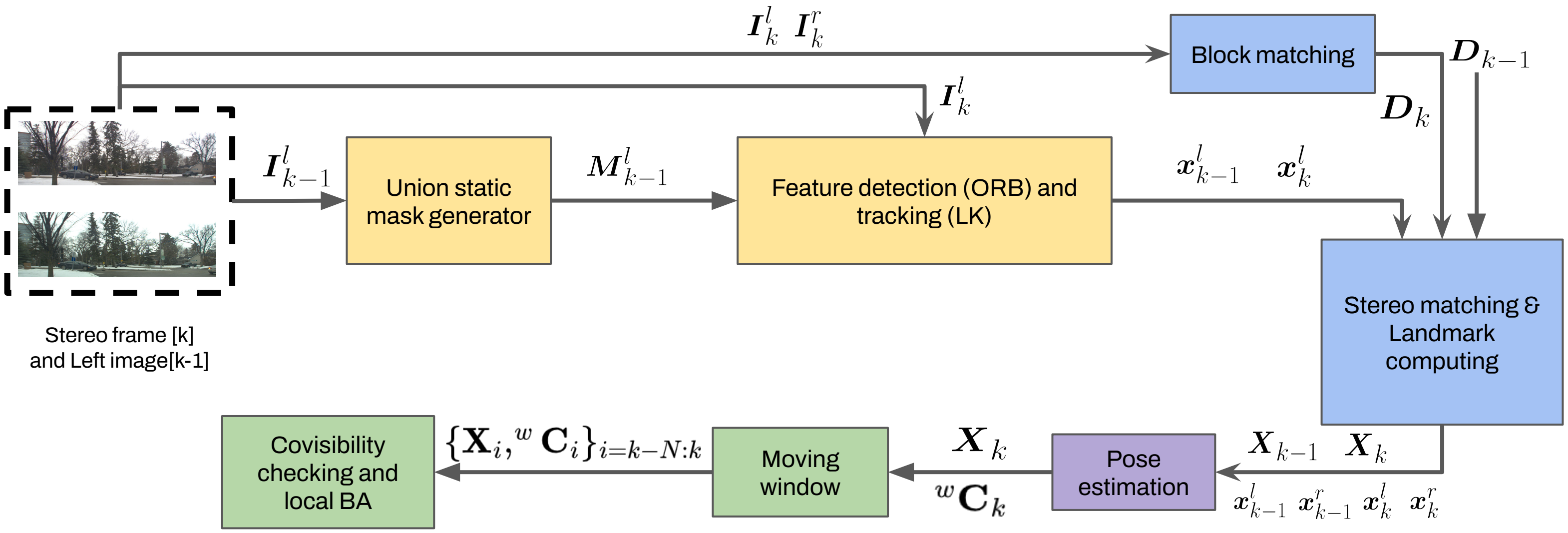
node LAB

4) Current pose C and landmarks in global frame are append to a moving window that consider N last frames.



SVO algorithm overview

4) Current pose C and landmarks in global frame are append to a moving window that consider N last frames. Covisibility checking ensures node connections.

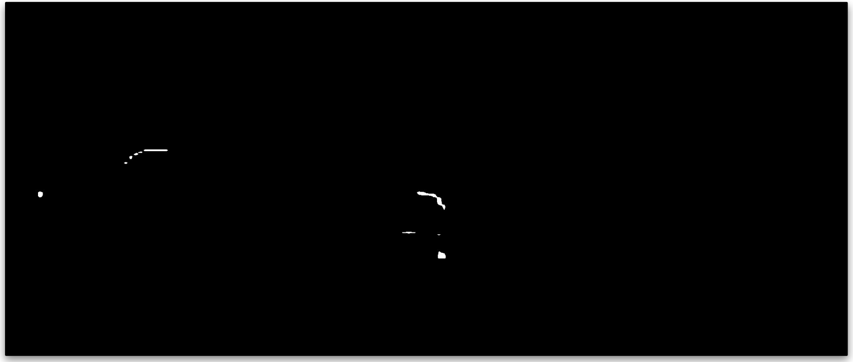


Urban union static mask

node LAB

- In urban canyons, we have certainty that there are static objects which cannot be moveable, thus avoiding movement discrimination in dynamic instance filtering.

RGB-D image



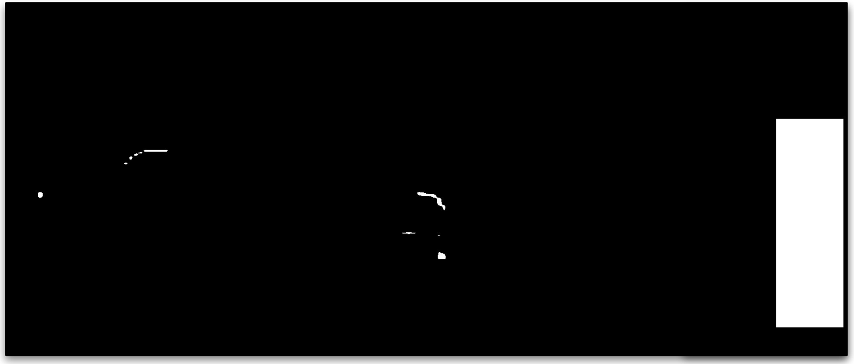
Light poles **U** *traffic lights* **U** *traffic signs* (from YOLACT[3])

[3] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, “YOLACT: real-time instance segmentation,” CoRR, vol. abs/1904.02689, 2019.

Urban union static mask

node LAB

RGB-D image



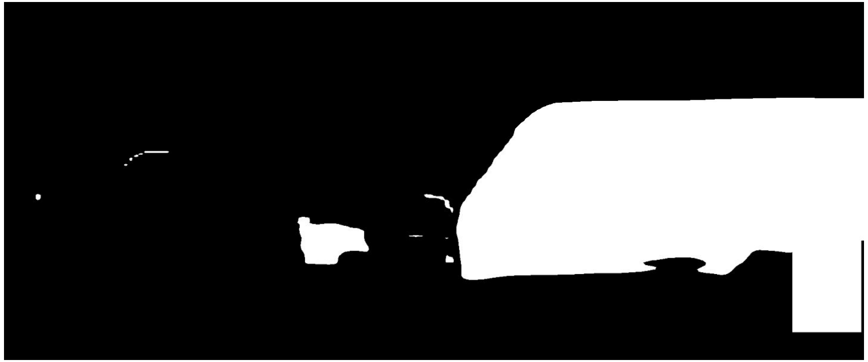
Light poles **U** *traffic lights* **U** *traffic signs*
U *tree trunk* (from YOLOv5[4])

[4] J. Solawetz, “What is YOLOv5? A Guide for Beginners.” 1 2023.

Urban union static mask

node LAB

RGB-D image



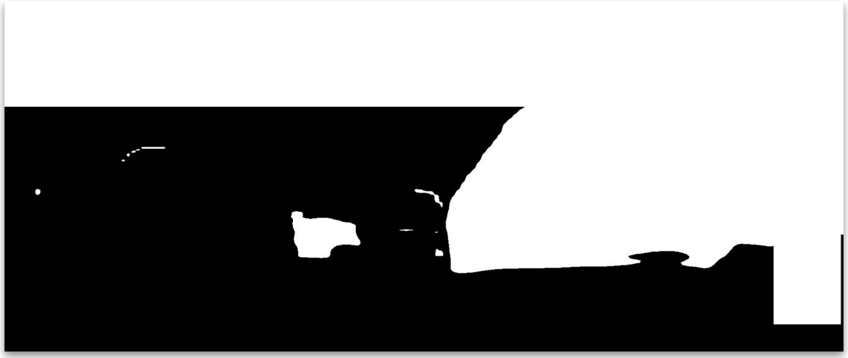
Light poles **U** *traffic lights* **U** *traffic signs* **U** *tree trunk* **U** *buildings* (from YOLACT[3])

Urban union static mask

node LAB

- A 30% upper image (tunnable depending on scene content) has been added to the mask to increase robustness against false positives in segmentation and detection.

RGB-D image



Light poles **U** *traffic lights* **U** *traffic signs* **U** *tree trunk* **U** *buildings* **U** *30%upper image*

Covisibility graph with static landmarks

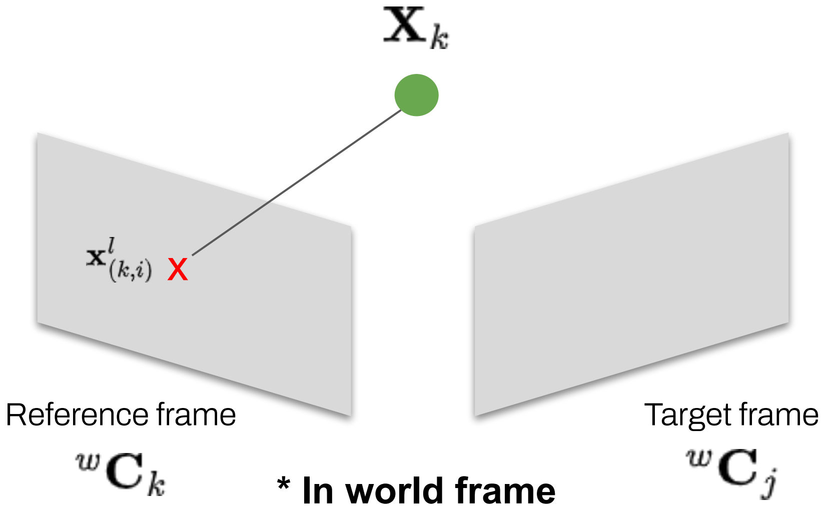
node LAB

- Took inspiration of ORB-SLAM [5] and Strasdat.et al [6] to adapt a covisibility graph to connect consecutive frames, map points and feature measurements.

Covisibility graph with static landmarks

node LAB

- Took inspiration of ORB-SLAM [5] and Strasdat.et al [6] to adapt a covisibility graph to connect consecutive frames, map points and feature measurements.



Algorithm 2: Covisibility check for one frame

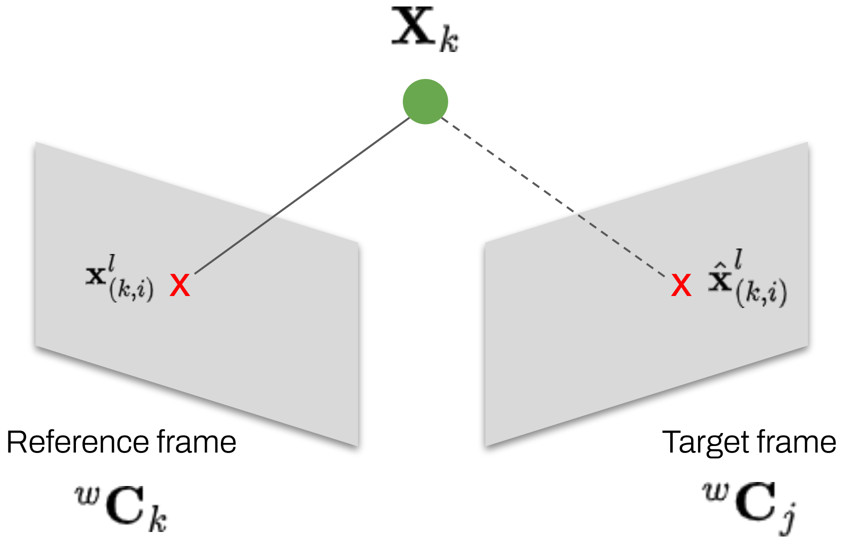
Initial : reference frame ${}^wC_k$, graph $\mathcal{C}$
Parameters: image dimensions (h, w)
$X_k = {}^w C_k X_k$;

[5] R. Mur-Artal, J. M. M. Montiel, and J. D. Tard ́os, “ORB-SLAM: a versatile and accurate monocular SLAM system,” CoRR, vol. abs/1502.00956, 2015.
[6] H. Strasdat, A. J. Davison, J. Montiel, and K. Konolige, “Double window optimisation for constant time visual slam,” in 2011 IEEE ICCV, 2011, pp. 2352–2359.

Covisibility graph with static landmarks

node LAB

- The static information is embedded in the graph

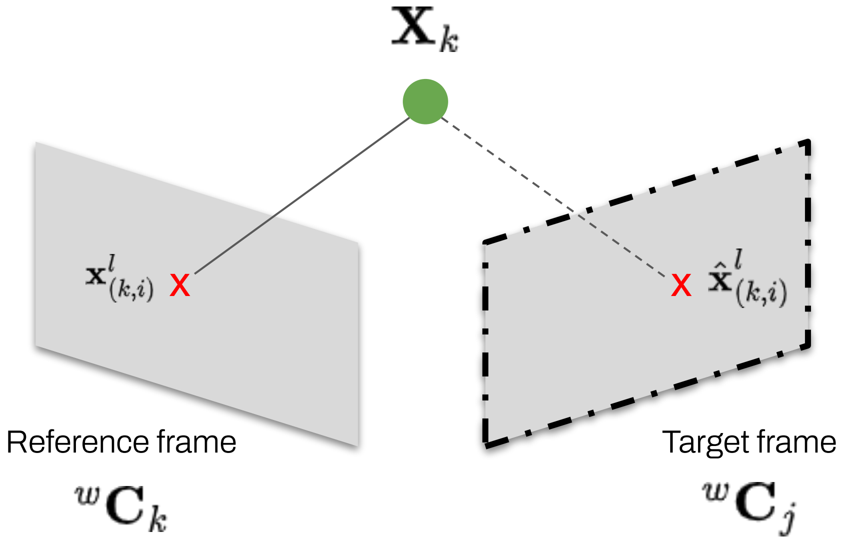


Algorithm 2: Covisibility check for one frame

Initial : reference frame ${}^wC_k$, graph $\mathcal{C}$
Parameters: image dimensions (h, w)
$X_k = {}^wC_k X_k$;
for *frame j in* $\mathcal{F} \leftarrow \mathcal{C}$ **do**
  if ${}^wC_j == {}^wC_k$ **then**
    pass;
  $\hat{x}^l_j = \pi(X_k, P^l, ({}^wC_j)^{-1})$;

IEEE ITSC 2023

Covisibility graph with static landmarks

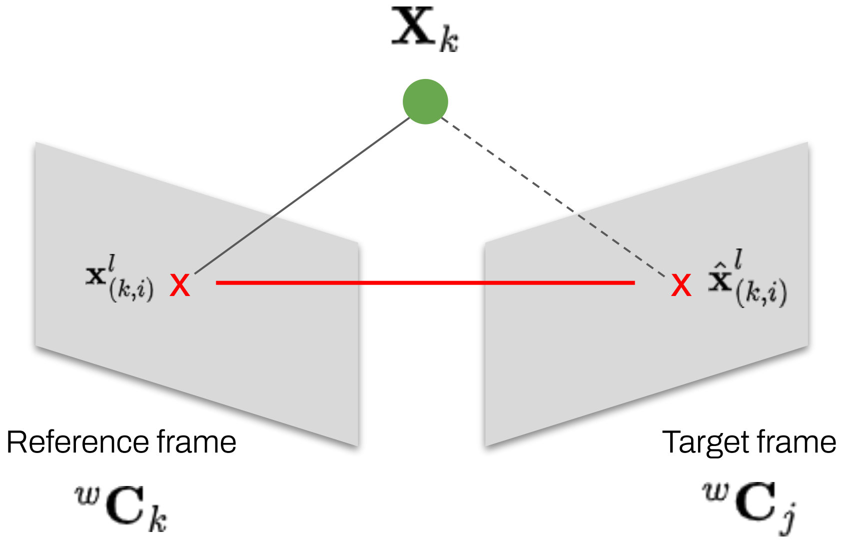


Algorithm 2: Covisibility check for one frame

Initial : reference frame ${}^wC_k$, graph $\mathcal{C}$
Parameters: image dimensions (h, w)
$X_k =^w C_k X_k$;
for *frame* j *in* $\mathcal{F} \leftarrow \mathcal{C}$ **do**
  if ${}^wC_j ==^w C_k$ **then**
    pass;
  $\hat{x}^l_j = \pi(X_k, P^l, ({}^wC_j)^{-1})$;
  for *point* i *in* $\hat{x}^l_j$ **do**
    if $\hat{x}^l_{(j,i)} \notin [h, w]$ **then**
      Reject $x^l_{k,i}, X_{k,i}$ in next steps;
  end

Covisibility graph with static landmarks

node LAB



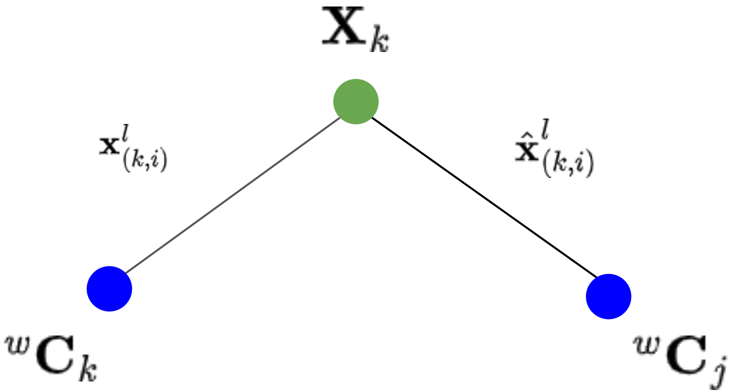
Algorithm 2: Covisibility check for one frame

Initial : reference frame ^wC_k, graph $\mathcal{C}$
Parameters: image dimensions (h, w)
$\boldsymbol{X}_k =^w C_k \boldsymbol{X}_k$;
for *frame* j *in* $\mathcal{F} \leftarrow \mathcal{C}$ **do**
  if $^wC_j ==^w C_k$ **then**
    pass;
  $\hat{x}^l_j = \pi(\boldsymbol{X}_k, \boldsymbol{P}^l, (^w\boldsymbol{C}_j)^{-1})$;
  for *point* i *in* $\hat{x}^l_j$ **do**
    if $\hat{x}^l_{(j,i)} \notin [h, w]$ **then**
      Reject $\boldsymbol{x}^l_{k,i}, \boldsymbol{X}_{k,i}$ in next steps;
  end
  $\mathcal{M} \leftarrow \mathcal{FM}(\boldsymbol{d}^l_k \leftarrow \boldsymbol{x}^l_k, \boldsymbol{d}^l_j \leftarrow \boldsymbol{x}^l_j)$;
  for (m', m) *in matches* $\mathcal{M}$ **do**
    Add observation $(^w\boldsymbol{C}_j, \boldsymbol{x}^l_{(j,m)})$ in $\boldsymbol{X}_{(k,m')}$;
  end
end

Covisibility graph with static landmarks

node LAB

Local bundle adjustment graph



extended to cover all the window poses and associated landmarks ...

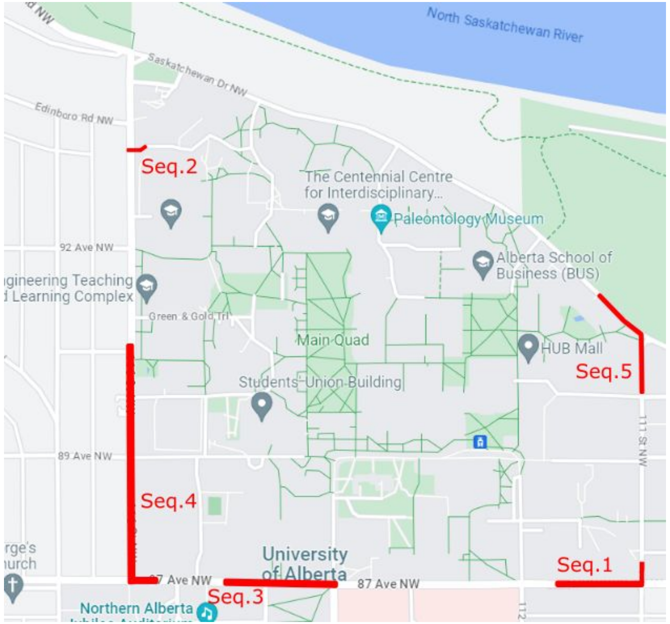
The graphs contains only static information.

Evaluation platform

node LAB

NODE Lab test car

Map view of test sequences

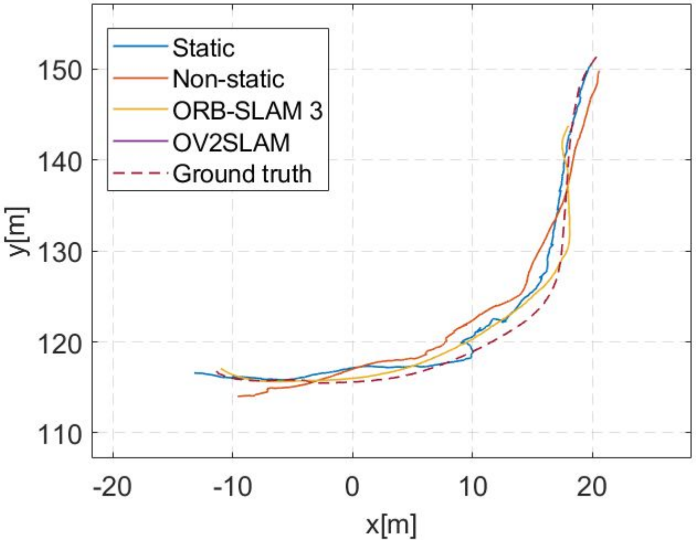


Trajectory results & comparison with SOTA

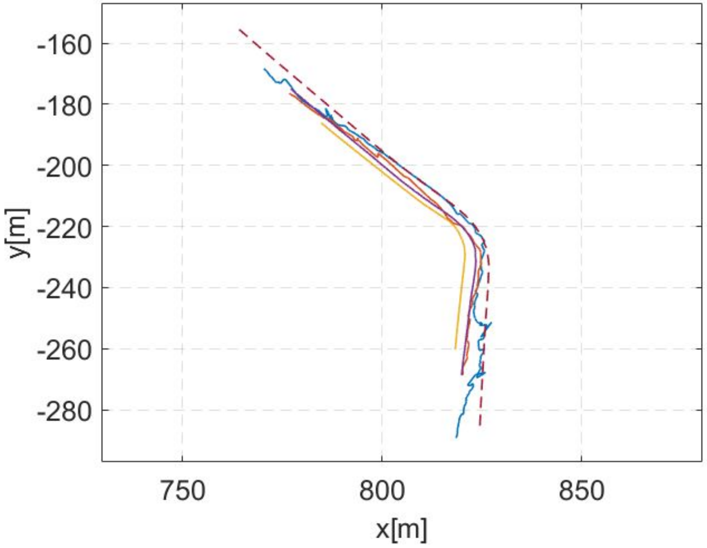
node LAB

The method handles scenarios under presence of dynamic objects (pedestrians and moving cars).

Sequence 2



Sequence 5



Quantitative results

node LAB

Seq.	Evaluation metric: ATE_T[m], ATE_R[rad]			
	Ours(static)	Ours(non-static)	ORB-SLAM 3	OV2SLAM
1	**5.165** / **0.455**	7.591 / 0.457	36.984 / 0.964	22.113 / 0.829
2	**1.975** / 0.356	2.544 / 0.288	2.822 / **0.090**	3.992 / 0.653
3	**6.577** / 0.517	8.377 / 0.562	24.867 / **0.106**	36.653 / 0.588
4	**10.349** / 0.506	12.402 / 0.336	25.052 / 0.143	16.579 / **0.140**
5	**6.959** / 0.357	10.863 / 0.474	17.72 / **0.123**	11.391 / 0.136

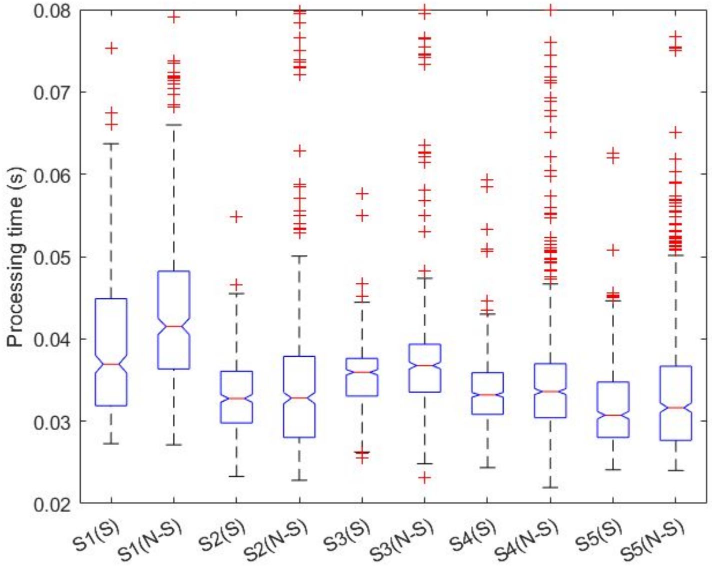
TABLE I: Comparison of different algorithms in dynamic sequences from NODE Lab dataset

*Best results are shown in **bold**

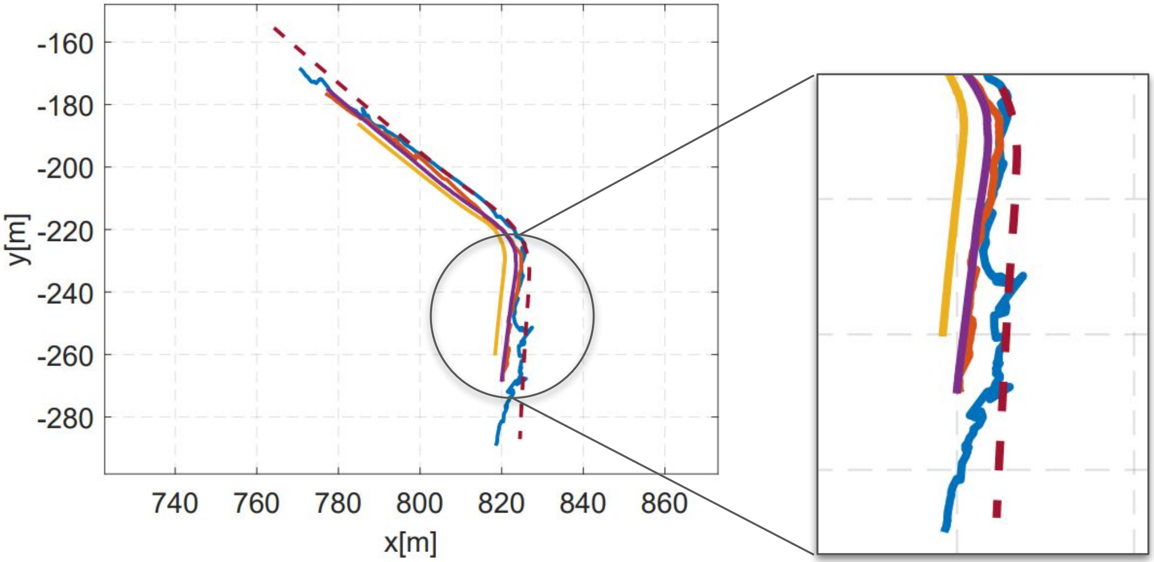
Processing time comparison

node LAB

The sparse set of static features reduces computational load in stages of stereo matching, triangulation and pose estimation in comparison to use the non-static set.



Issues and future works



Temporal instability on instance segmentation or object detection due to **occlusion**, **fast movement** or **incapability to process far objects** injects noise in estimation.

Conclusions

node LAB

- A semantic-aware stereo visual odometry has been presented which identifies street objects to include them in static ROI for reliable feature extraction.

- Static features, landmarks and poses were wrapped inside a local bundle adjustment optimization. The refinement BA is benefited by using a reduced set of keypoints for faster and more accurate results.

- The effectiveness of the method has been demonstrated through extended evaluation on urban canyons (University of Alberta) against other methods.

Funding Agencies and Acknowledgements

NSERC
CRSNG

Thank You!

NODE Lab
e-mail: marceloj@ualberta.ca, npbhatt@uwaterloo.ca and ehashemi@ualberta.ca