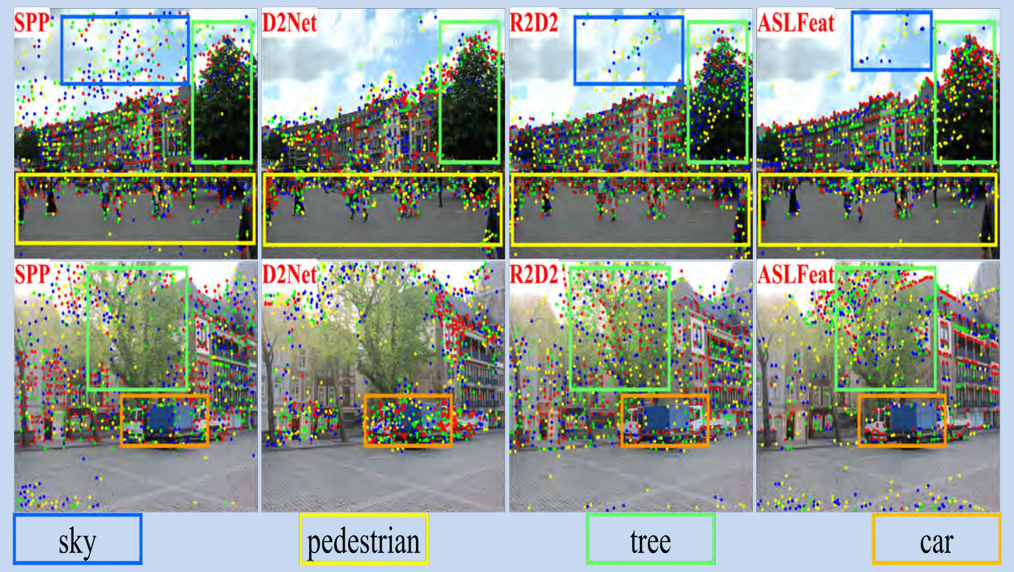
SFD2: Semantic-guided Feature Detection and Description

UNIVERSITY OF CAMBRIDGE

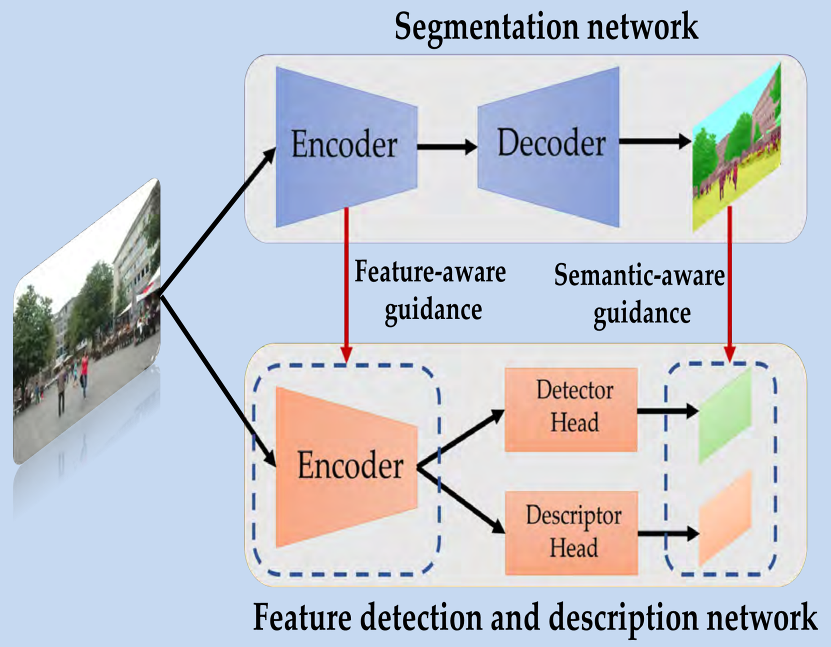
Fei Xue, Ignas Budvytis, Roberto Cipolla

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- **Problem**
  - Learning robust features for long-term visual localization
- **Local features only**
  - Focusing mainly on *local* reliability
  - Performing exhaustive detection
  - Sensitive to viewpoint, illumination, season changes and dynamic objects
- **Local features + explicit semantics**
  - Requiring accurate semantic labels at test time
  - Fragile to segmentation errors especially for night images
- **Motivation**
  - Implicitly embedding semantics into local features
  - No need of ground-truth semantics for training
    - Learning from off-the-shelf segment network in the training
    - Semantic–aware and feature-aware guidance
  - No explicit semantic labels required at test time



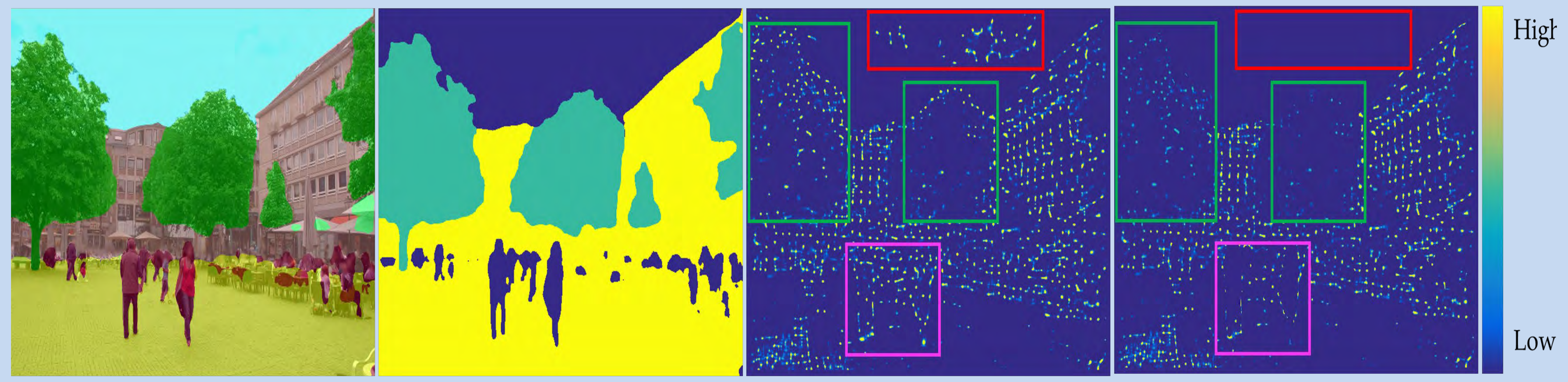
Exhaustive detection of prior local features



Our framework

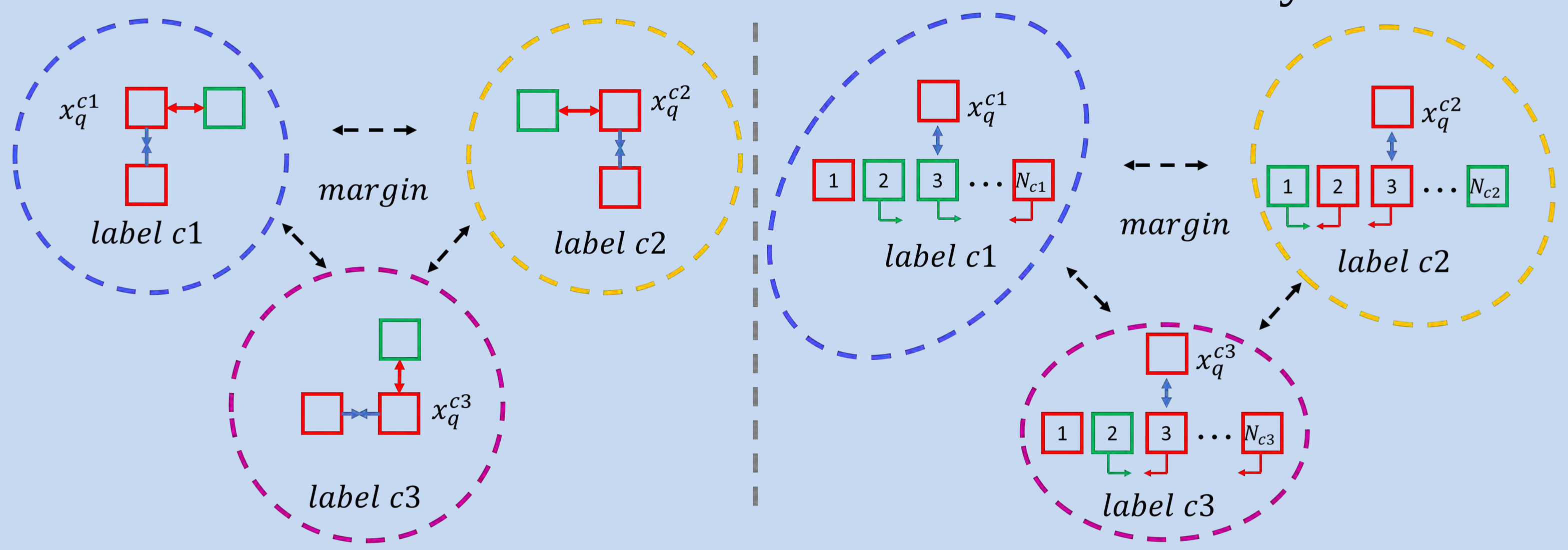
[1] LBR, Xue, et al., CVPR 2022
[2] Suerpoint, DeTone et al., CVPRW 2018
[3] R2D2, Revaud et al., NeurIPS 2019
[4] SuperGlue, Sarlin et al, CVPR 2020
[5] ASLFeat, Luo et al., CVPR 2020
[6] ConvNeXt, Liu et al, CVPR 2022

- **Approach**
- **Semantic-aware guidance - detection**
  - Redefined reliability = local reliability ⊗ stability
  - Favoring features from robust areas, e.g., buildings
  - Suppressing features from unstable areas, e.g., trees, pedestrians



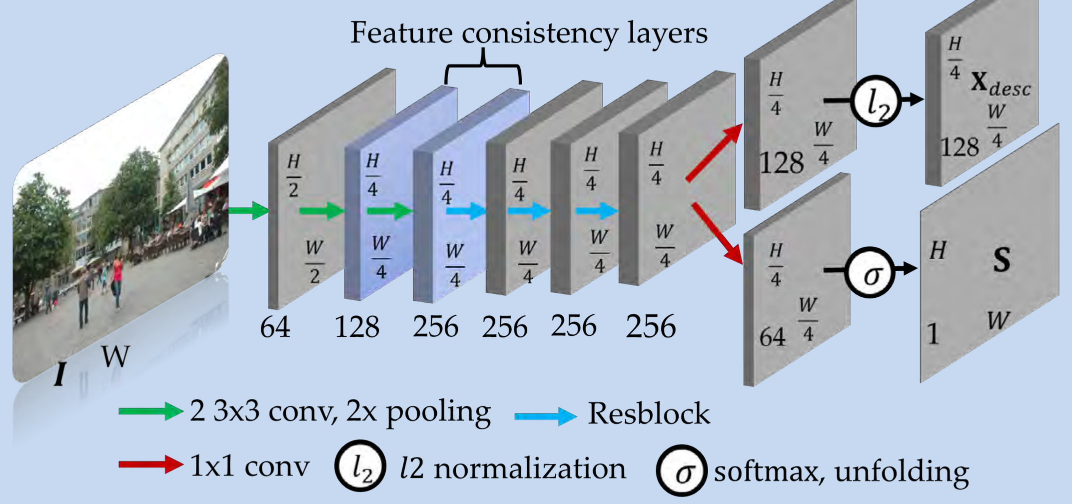
Semantic mask | Stability map | Reliability map | Final reliability map

- **Semantic-aware guidance - description**
  - Inter-class loss – features with same labels should be close, otherwise far
  - Intra-class loss – ranking positive/negative keypoints with same labels
  - Flexible to classes with various discriminative ability



pull force · intra-class push force · inter-class push force

- **Feature-aware guidance**
  - Feature consistency between encoded features (ConvNeXt for supervision)
  - Style transfer learning of night images
  - More effective at embedding high-level semantic information



- **Results**
- **Robust semantic-aware detection and matching**
  - 1k keypoints with reliability from high to low (red, green, blue, yellow)

Superpoint | R2D2 | ASLFeat | **Ours**

More on buildings
Fewer on trees
Robust to night images

Robust to viewpoint changes

Robust to illumination/season changes

- **More accurate localization accuracy on Aachen dataset (day/night)**

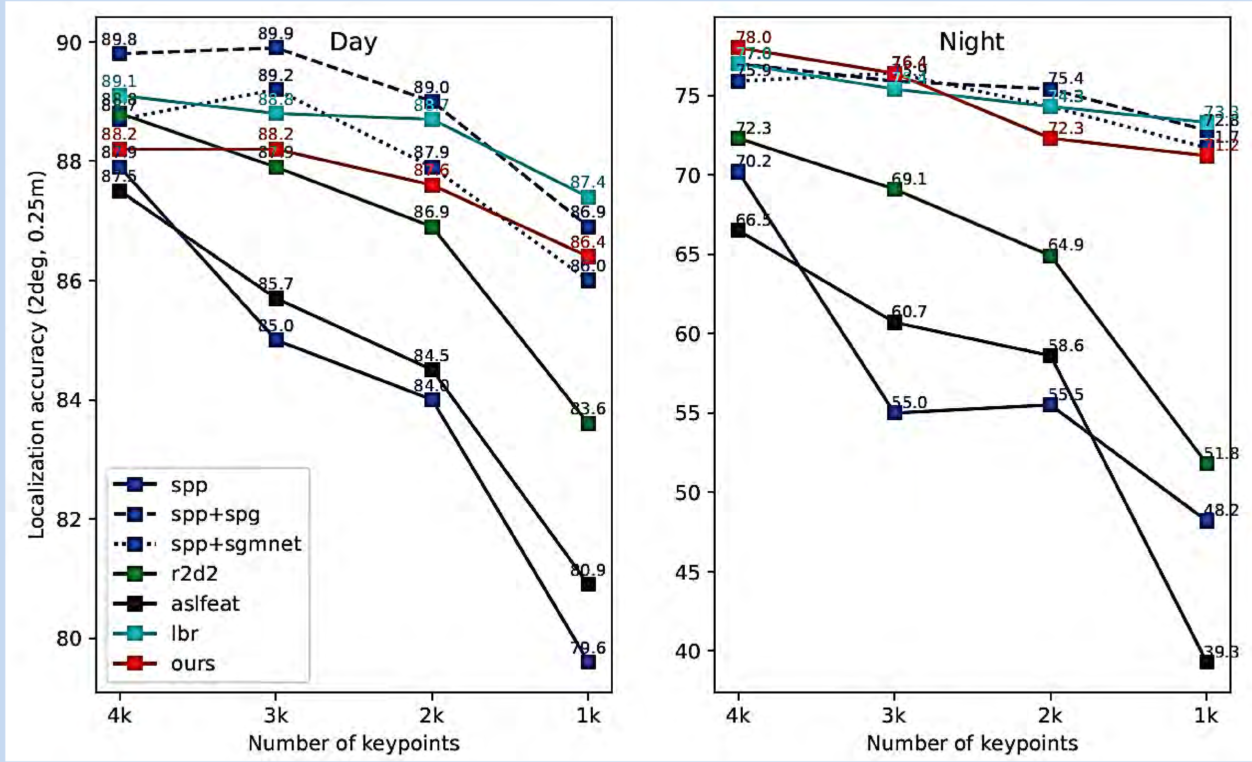
	$2°, 0.25m/5°, 0.5m/10°, 5m$	
LBR	88.3 / 95.6 / **98.8**	84.7 / 93.9 / **100.0**
R2D2	N/A	84.7 / 90.8 / 96.9
ASLFeat	N/A	81.6 / 87.8 / **100.0**
SPP+SPG	**89.6** / 95.4 / **98.8**	86.7 / 93.9 / **100.0**
Ours	88.2 / **96.0** / 98.7	**87.8** / **94.9** / **100.0**

Better than methods with explicit labels

Better than prior local features

Close to advanced matchers

- **Robust to keypoint reduction and faster**



Localization accuracy with different number of keypoints

- 2.2x faster than R2D2
- 3.4x faster than ASLFeat
- 4.4x faster than SPP+SPG

LBR	39.3
R2D2	72.4
ASLFeat	112.3
SPP+SPG	146.5
Our	**33.2**

Test time of 1024 keypoints (ms)