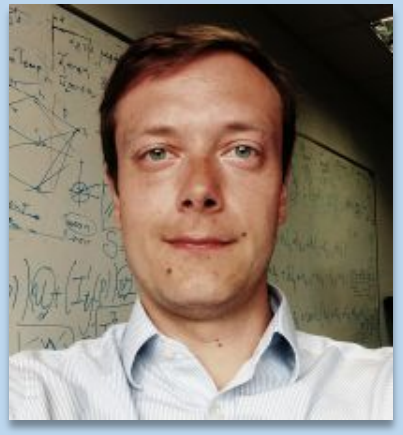
UNIVERSITY OF CAMBRIDGE

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

SFD2: Semantic-guided Feature Detection and Description

Fei Xue

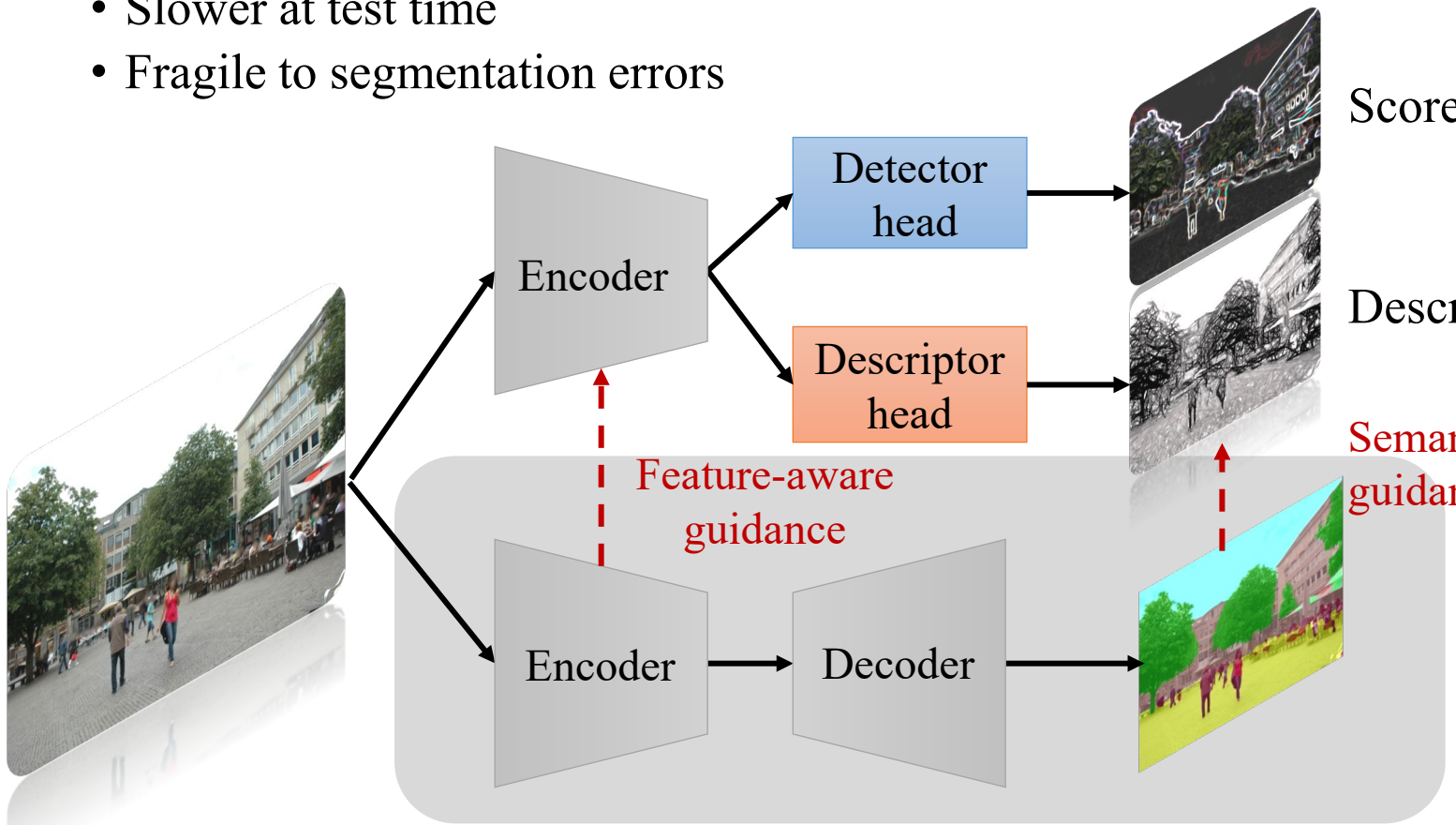


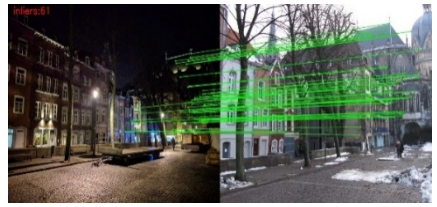
Ignas Budvytis

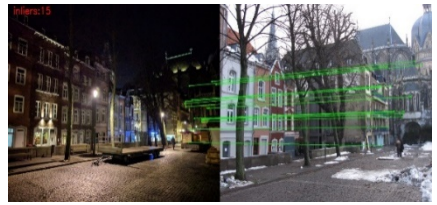
Roberto Cipolla

Preview of SFD2

- **Feature detection & description**
  - Long-term localization
  - Sensitive to season changes
- **Explicit semantics**
  - Slower at test time
  - Fragile to segmentation errors
- **Implicit semantic embedding**
  - Semantic-aware guidance
  - Feature-aware guidance
- **No explicit semantics at test time**
  - Faster
  - Less fragile to segmentation errors
- **Robust to appearance changes**



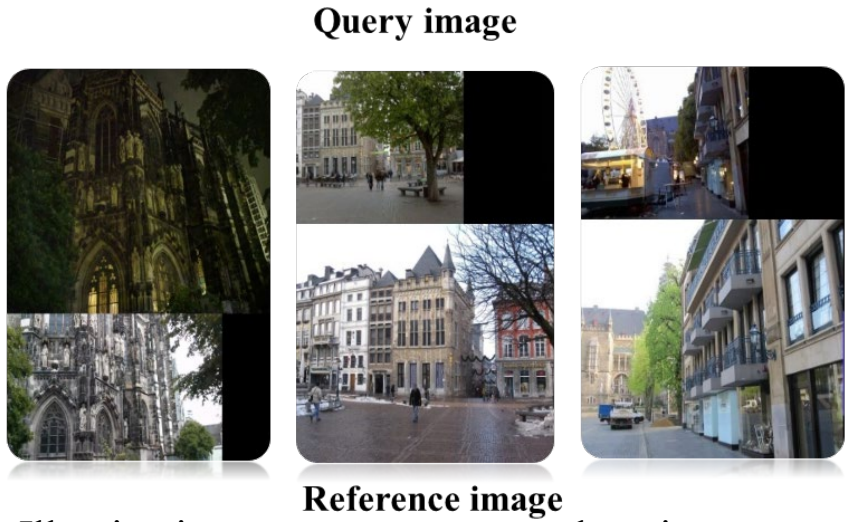

Ours

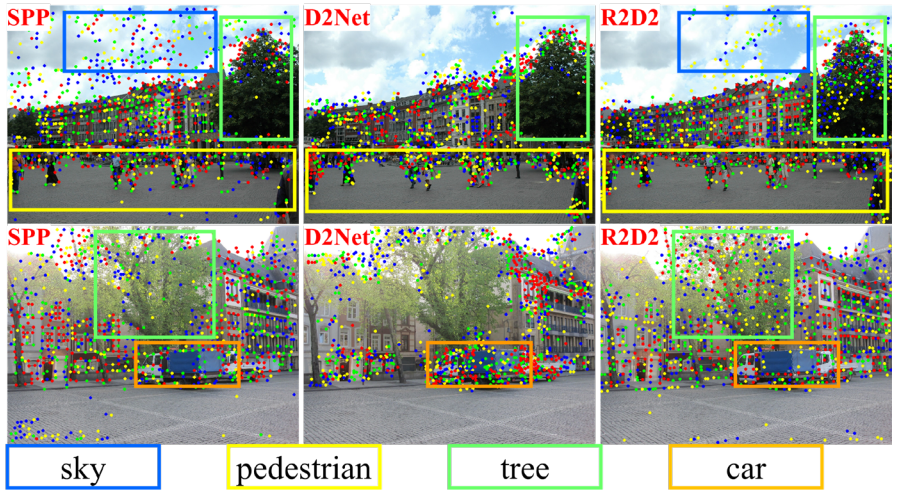

Superpoint


R2D2

Local features are key to localization

- **Challenges of long-term localization**
  - Large viewpoint changes
  - Severe illumination and seasonal changes
  - Dynamic objects
- **Prior features are local**
  - Indiscriminative detection
  - Sensitive to above challenges





Many useless keypoints from **sky**, **trees**, **cars**

Superpoint

R2D2

Localization errors

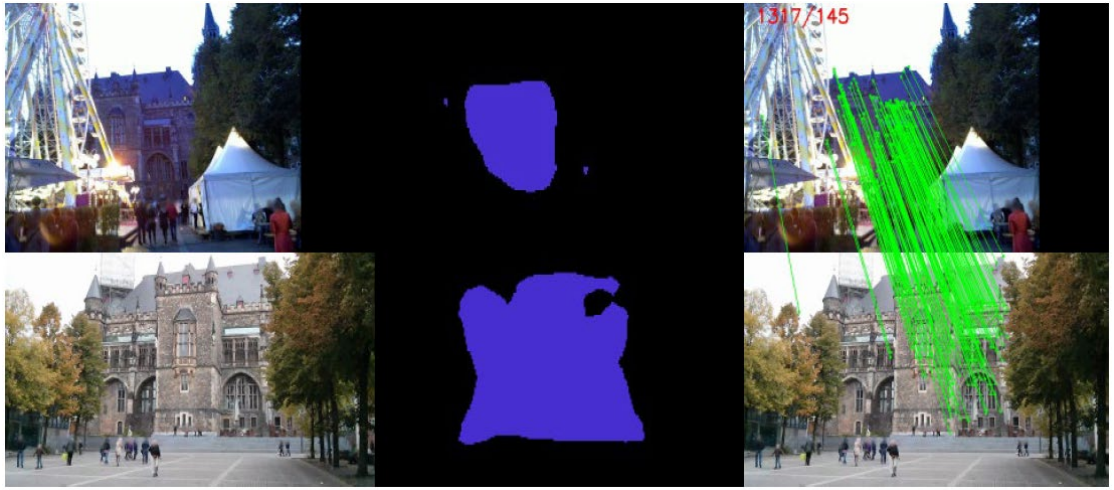
[1] SPP: DeTone et al., CVPRW 2018
[2] D2Net: Dusmanu et al., CVPR 2019
[3] R2D2: Revaud et al, NeurIPS 2019

Local features are key to localization

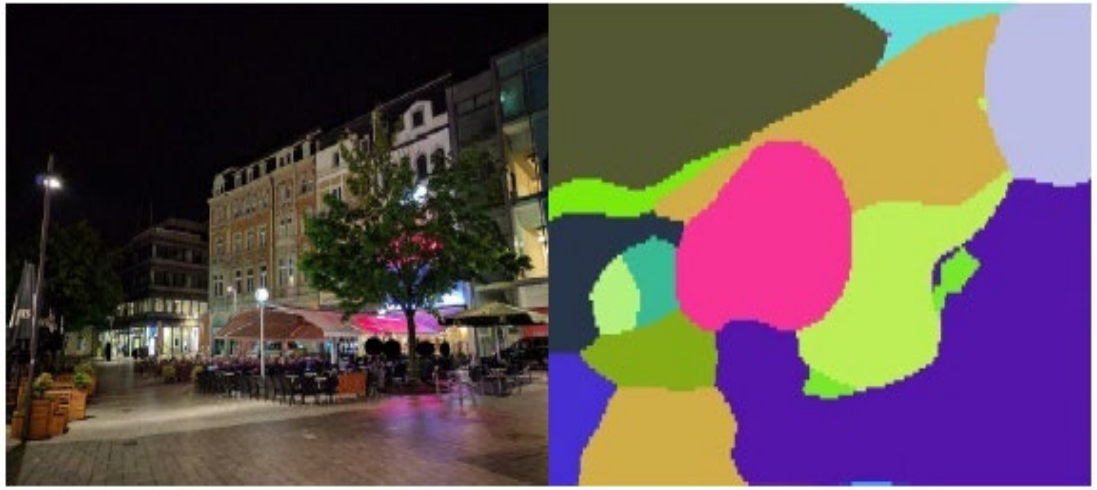
- **Semantic-aware localization**
  - More robust to appearance changes
  - Need explicit semantic labels at test time
  - Fragile to wrong segmentation results

Semantic mask

Areas with different stability



Semantics and can be used for matching



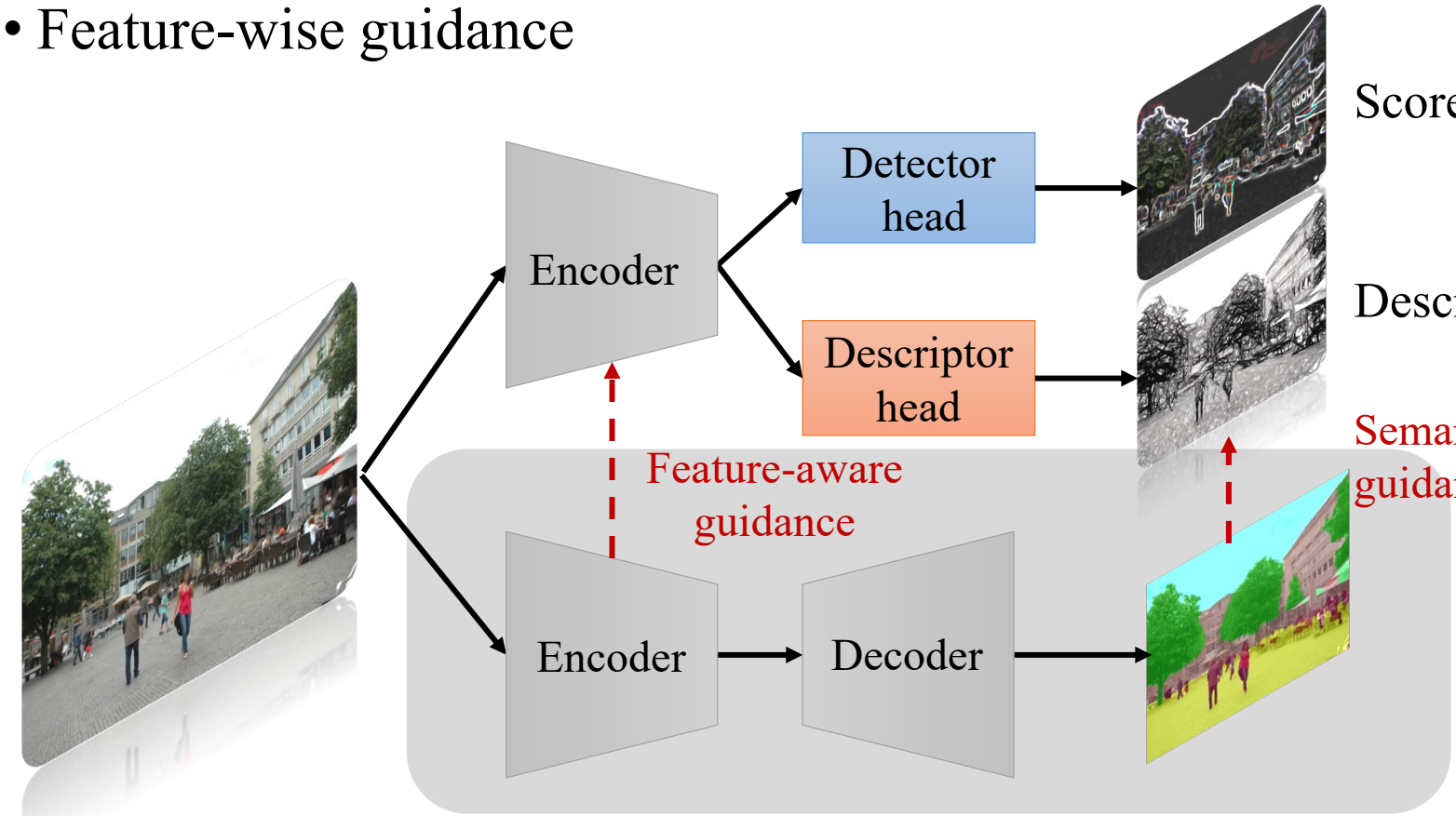
Usage of explicit semantic labels are fragile to segmentation errors

[1] LBR: Xue et al., CVPRW 2022

Implicit semantic embedding

- **Implicit semantic embedding**
  - Learning from segmentation networks
  - Semantic-wise guidance
  - Feature-wise guidance

- Semantics are embedded into feature network
- No need of explicit semantics at test time



End-to-end feature detection & description

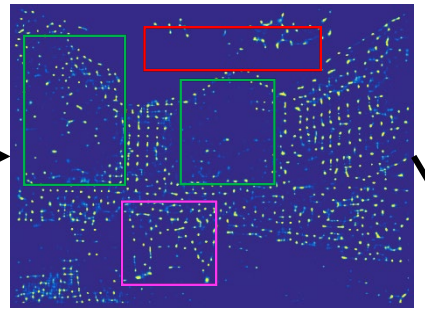
Off-the-shelf segmentation network

Semantic-aware guidance – detection

- **Local reliability**

Detector head

Tree, pedestrian, sky→useless for long-term localization



Local reliability map S_l

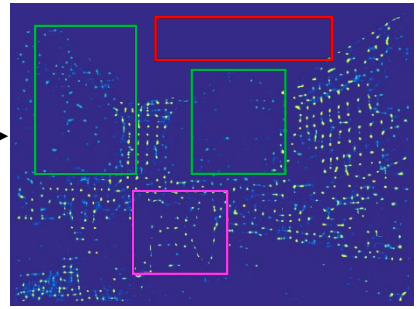
Stable features - retained
Unstable features - suppressed

- **Global stability**

Segmentation network

Semantic mask

Stability map S_g



$S = S_l \odot S_g$

Final reliability

Assign each class a stability value

Category	Class	Stability value
Volatile	water, etc	0.1
Dynamic	pedestrian, etc	0.1
Short-term	tree, etc	0.5
Long-term	building, etc	1.0

Semantic-aware guidance – description

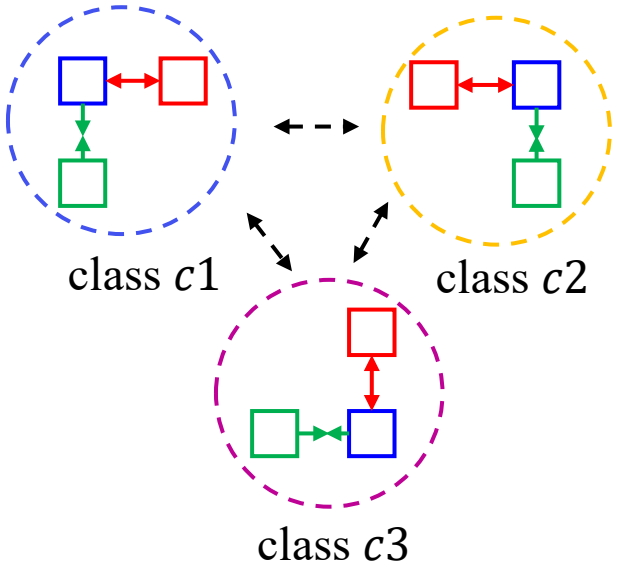
- **Inter-class discrimination**
  - Min. dist. of same classes
  - Max. dist. of different classes

$$L_{inter} = \frac{1}{N}\sum(\|x_i^{c1} - x_j^{c1}\|_2 - \|x_i^{c1} - x_k^{c2}\|_2 + m)$$

Descriptor head

Descriptor map X

Query Positive Negative



- **Intra-class discrimination**

Triplet loss
Min. dist. of pos. samples
Max. dist. of neg. samples

Ranking loss
Min. ranks of pos. samples
Max. ranks of neg. samples

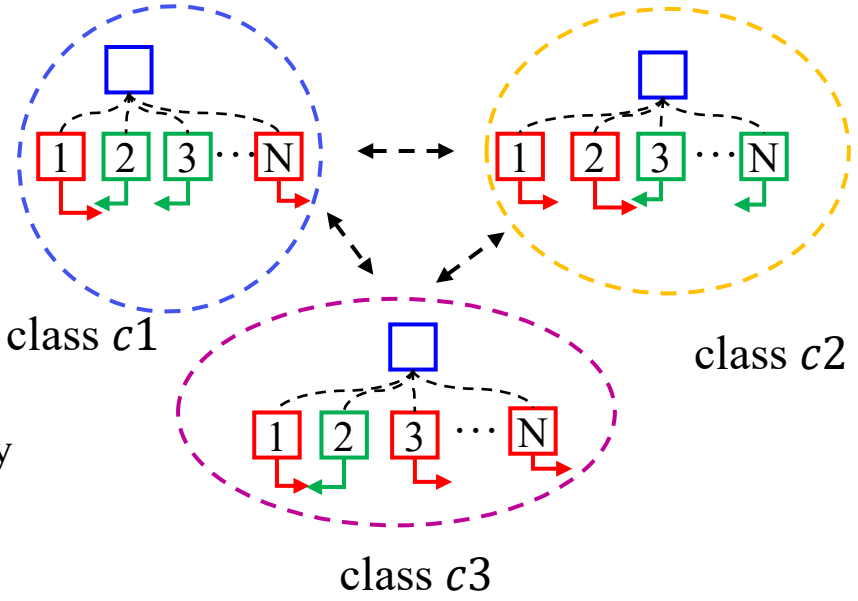
$$L_{intra} = \frac{1}{NC}\sum\sum f_{ranking}(x_i^c)$$

Not retain inner discriminative ability

Low discriminative ability

High discriminative ability

Retain inner discriminative ability



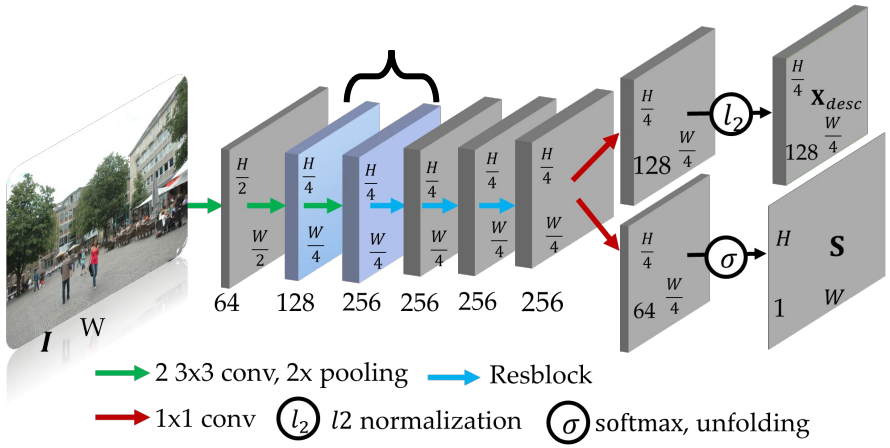
[1] Ranking loss: He et al., CVPR 2018

Feature-aware guidance

- **High-level semantics are not easy to learn**
- **Feature-consistency loss**

$$L_{feat} = \frac{1}{2}\sum_{i} |X_i - X_i^{seg}|$$

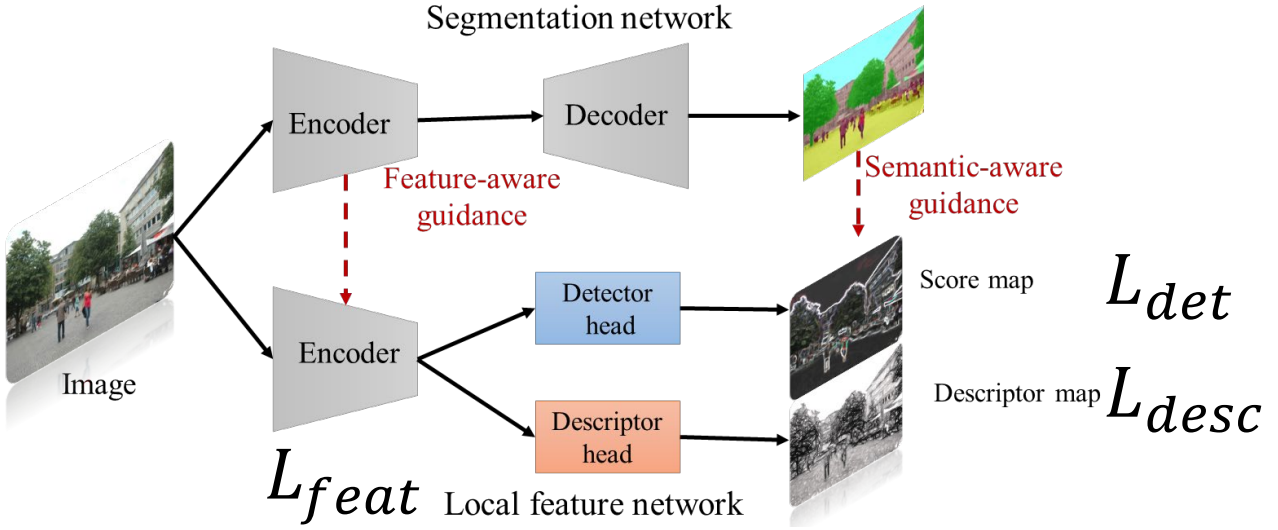
Consistent with segmentation networks



Architecture of our network

- **Total loss**
  - Semantic-aware detection loss
  - Semantic-aware description loss
  - Feature-consistency loss
  - Superpoint → local reliability
  - ConvXt → semantic labels

$$L_{total} = \alpha_{det}L_{det} + \alpha_{desc}(L_{inter}+L_{intra}) + \alpha_{feat}L_{feat}$$

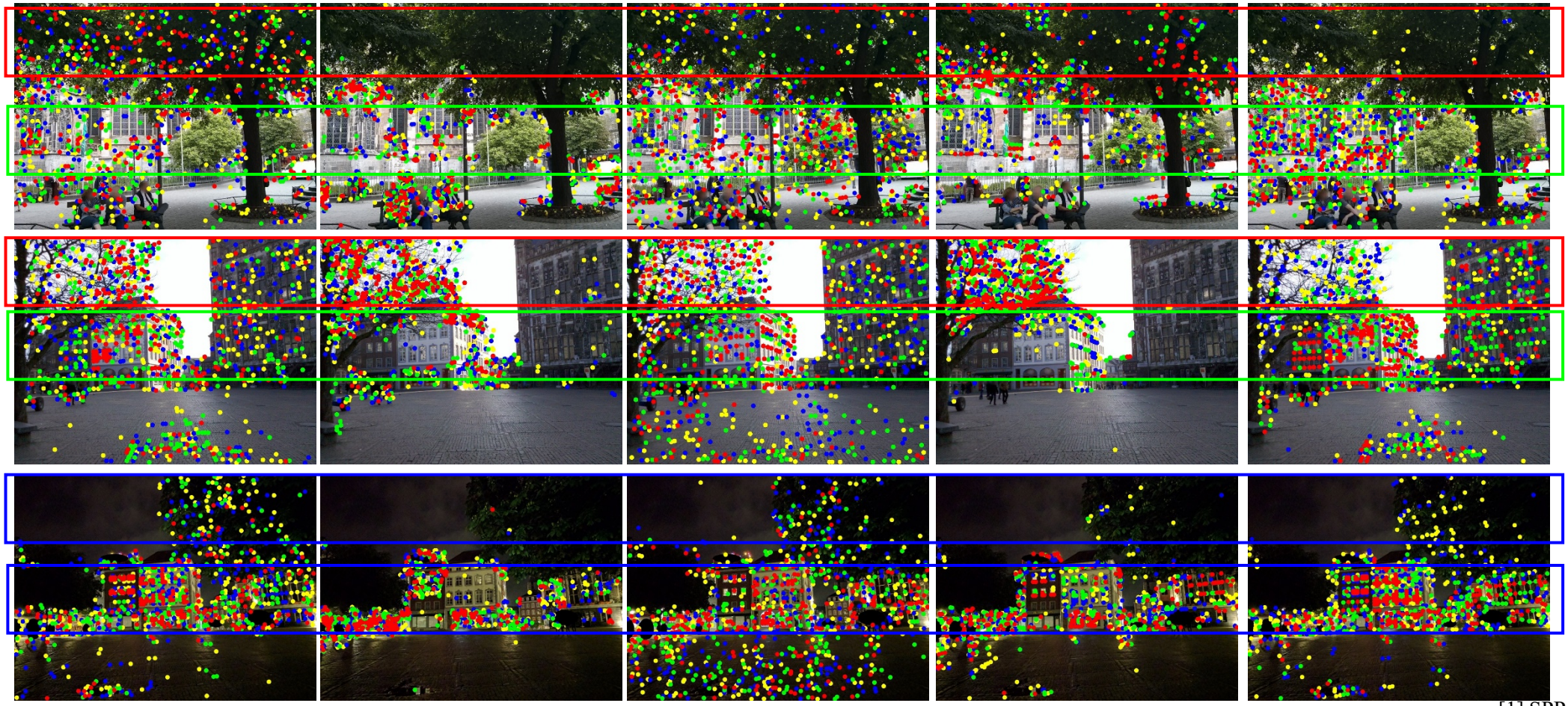


[1] SPP: DeTone et al., CVPRW 2018
[2] ConvXt: Liu et al., CVPR 2022

Experiments-detection

- **1000 keypoints (top 1-250, 250-500, 500-750, 750-1000)**

SuperPoint D2Net R2D2 ASLFeat Ours



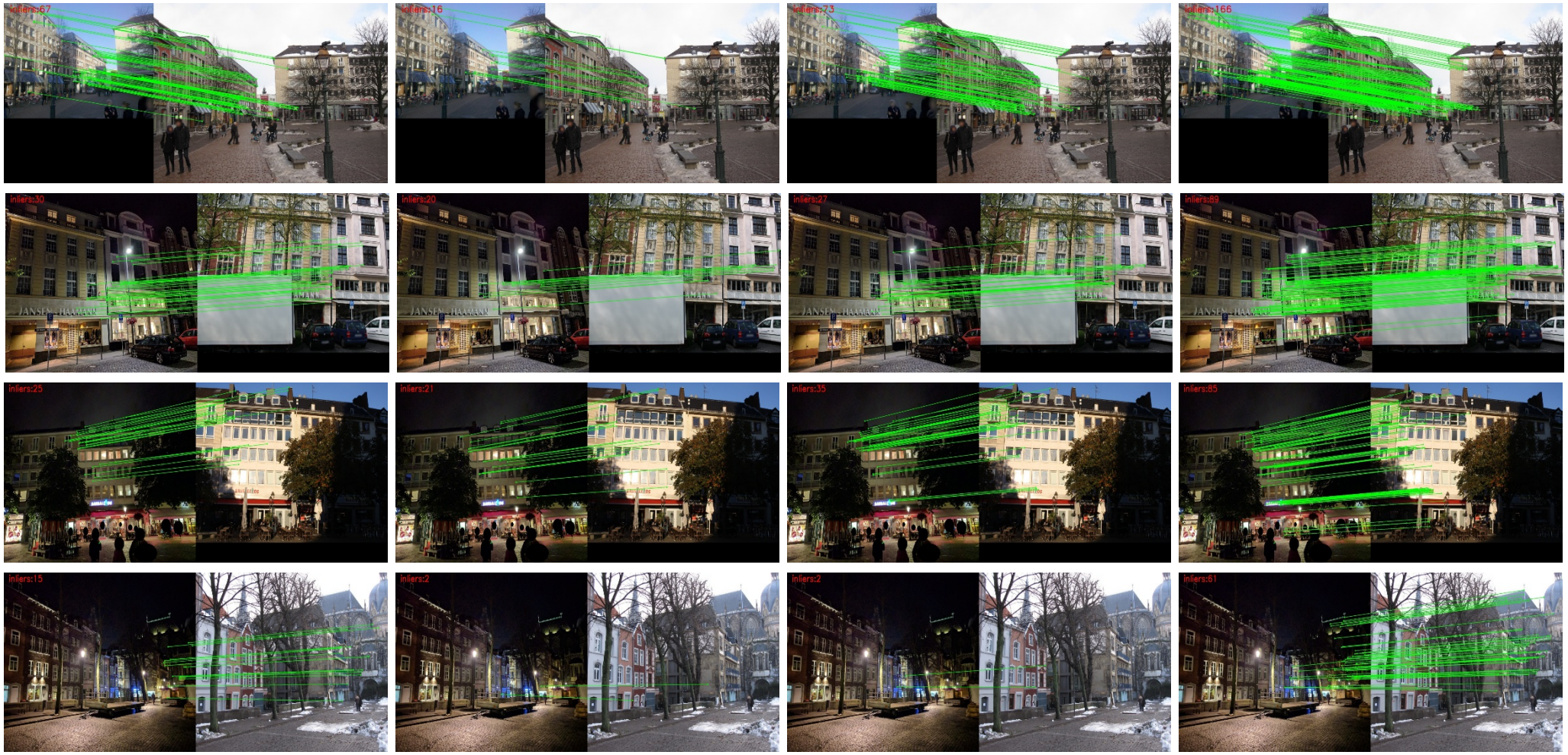
- Fewer on trees
- More on buildings
- Robust to night images

[1] SPP: DeTone et al., CVPRW 2018
[2] D2Net: Dusmanu et al., CVPR 2019
[3] R2D2: Revaud et al, NeurIPS 2019
[4] ASLFeat: Luo et al., CVPR 2020

Experiments

- **Inliers between query and reference images**

SuperPoint | R2D2 | ASLFeat | Ours



- More robust to viewpoint changes
- More robust to dynamic objects
- More robust to illumination changes
- More robust to seasonal changes

More accurate for long-term localization

- Aachen Day Night and RobotCar Seasons datasets
- Semantic-aware methods (S) (e.g., LBR, SSM)
- Local features (L) (e.g., SuperPoint, R2D2, ASLFeat)
- Advanced matchers (M) (e.g., SuperGlue, SGMNet)

Localization at error thresholds of $0.25m, 2^o$ /$0.5m, 5^o$/$5m$ 10^o

Best and second-best results are highlighted

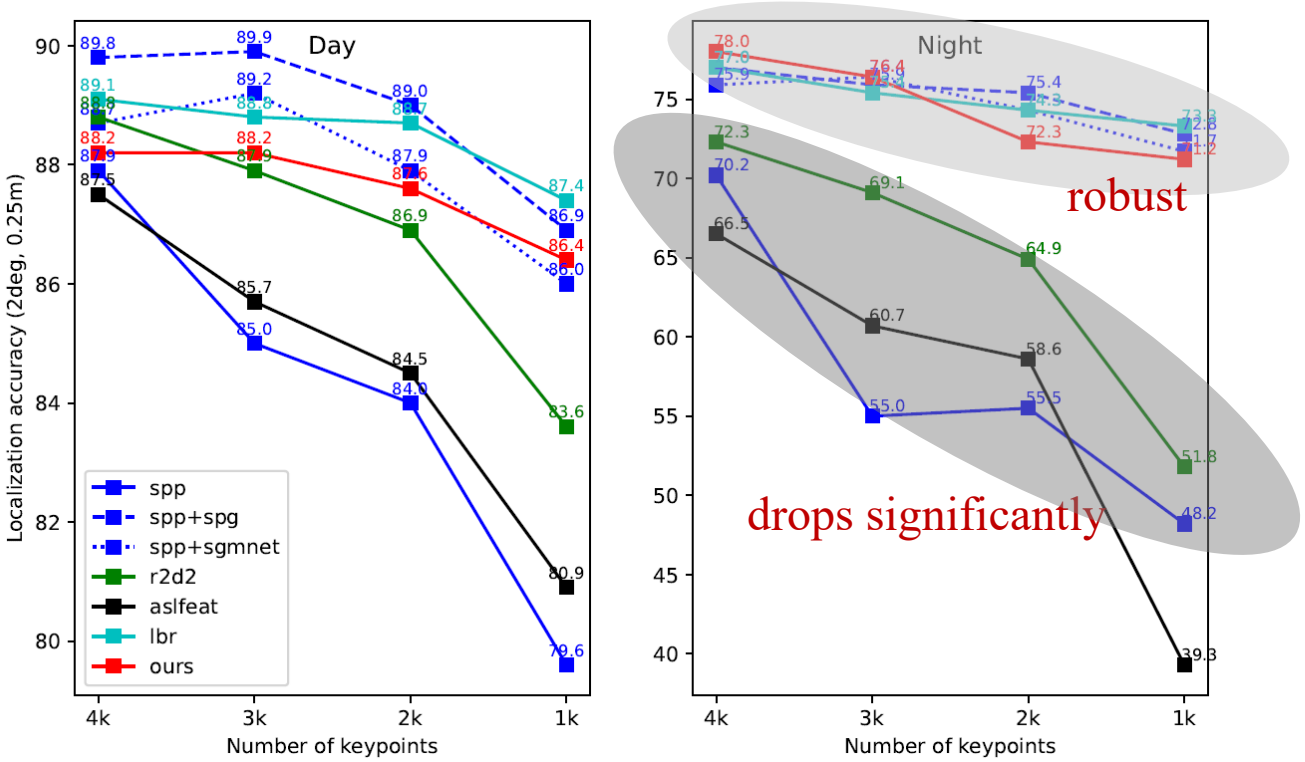
Group	Method	Day	Night	Day	Night	Night-rain
S	SSM	71.8 / 91.5 / 96.8	58.2 / 76.5 / 90.8	54.5 / 81.6 / 96.7	10.0 / 23.7 / 45.4	14.5 / 33.2 / 47.5
	LBR	**88.3** / 95.6 / **98.8**	84.7 / 93.9 / **100.0**	56.7 / **81.7** / **98.2**	24.9 / 62.3 / 86.1	**47.5** / **73.4** / 90.0
	Ours	88.2 / **96.0** / 98.7	**87.8** / **94.9** / **100.0**	**56.9** / 81.6 / 97.4	**27.6** / **66.2** / **90.2**	43.0 / 71.1 / 90.0
L	Superpoint	80.5 / 87.4 / 94.2	42.9 / 62.2 / 76.5	56.5 / 81.5 / 97.1	16.9 / 41.6 / 71.5	22.0 / 45.0 / 68.0
	R2D2	N/A	76.5 / 90.8 / **100.0**	**57.4** / **81.9** / **97.9**	18.3 / 43.4 / 67.8	29.1 / 50.2 / 68.2
	Ours	**88.2** / **96.0** / **98.7**	**87.8** / **94.9** / **100.0**	56.9 / 81.6 / 97.4	**27.6** / **66.2** / **90.2**	**43.0** / **71.1** / **90.0**
M	SuperGlue	**89.6** / 95.4 / **98.8**	86.7 / 93.9 / **100.0**	**56.9** / **81.7** / **98.1**	24.2 / 62.6 / 87.4	42.3 / 69.3 / **90.2**
	SGMNet	86.8 / 94.2 / 97.7	83.7 / 91.8 / 99.0	N/A	N/A	N/A
	Ours	88.2 / **96.0** / 98.7	**87.8** / **94.9** / **100.0**	**56.9** / 81.6 / 97.4	**27.6** / **66.2** / **90.2**	**43.0** / **71.1** / 90.0

- Better
- Significantly better
- Close

[1] Aachen: Sattler et al., CVPR 2018
[2] SSM: Shi et al., ICIP 2019
[3] LBR: Xue et al., CVPR 2022
[4] SuperGlue: Sarlin et al., CVPR 2020
[5] SGMNet: Chen et al., ICCV 2021

Robust to keypoints changes & faster

- **Performance against #kpts (4k, 3k, 2k, 1k)**



Localization accuracy on Aachen at error thresholds of $0.5m, 5^o$

- **Running time**
  - Much faster than R2D2 and SuperGlue
  - Slower but more accurate than Superpoint
  - A good trade-off between accuracy and efficiency

Method	Time (ms)
LBR	39.3
SuperPoint	13.1
R2D2	72.4
SuperGlue	159.6
Ours	33.2

Running time on RTX 3090

Summary and future work

- **Summary**
  - Embedding semantics into local features implicitly
  - Semantic-aware and feature-aware guidance
  - More accurate and robust than prior competitors

- **Future work**
  - Semantic labels are based-on ADE20k → Learning semantic labels automatically
  - Current framework is designed for outdoor localization → A general model for both indoor and outdoor scenes