Optimal Transport Aggregation for Visual Place Recognition

Supplementary Material

6. Evaluation on Nordland

In this paper, we follow the evaluation pipeline used in MixVPR [2] and GSV-Cities [1]. This pipeline considers a retrieved image in Nordland as correct if it lies at less than two frames from the query image.

However, other works, like EigenPlaces [6], consider a different evaluation strategy for the NordLand dataset. They consider correct all images that are within ten frames from the query image. This setup results in higher recalls. Table 8 show results of our method and the previous state-of-the-art on Nordland using this evaluation pipeline. Our method also outperforms by a large margin the previous state-of-the-art.

7. Results of SuperGlobal

SuperGlobal [50] proposes three new modules that improve the recall of a model. Among these modules, Scale-GeM performs multi-scale aggregation, improving the retrieval when landmarks may appear at different scales. Multi-scale, however, is detrimental for VPR, as this task requires finding the closest viewpoint, for which landmark scale is a key clue. Due to this, as shown in Table 9, in this setup, SuperGlobal metrics are worse than GeM. They also propose to tune the p parameter of GeM, which may provide increased performance. However, this hyperparameter optimization is out of the scope of our paper.

8. Benchmark Description

MSLS Validation and Challenge: A large dataset of dashcam images in urban scenarios. It comprises a wide variation of cities, continents, season, and time of the day. Most of the images are forward facing. The challenge has closed labels and provides an online platform to evaluate models, reducing the saturation of the performance.

Nordland: Images captured from the front of a train traversing Norway. Query images are captured in summer and reference ones are from winter. Is a challenging benchmark given the high similarity of the images.

Pittsburgh-250k: A collection of urban Google Street View images featuring large viewpoint changes. For every place, it contains multiple images at different angles obtained from the same panoramic image.

SPED: It is comprised of CCTV images at different times. Therefore, it exhibits great time shift while keeping the exact same viewpoint of the places.

Method	Nordland		
	R@1	R@5	R@10
MixVPR	76.2	86.9	90.3
EigenPlaces	71.2	83.8	88.1
DINOv2 SALAD	**86.6**	**94.0**	**95.9**

Table 8. **NordLand evaluation.**

		MSLS Val		NordLand	
Method	Desc. size	R@1	R@5	R@1	R@5
ResNet GeM	1024	78.2	86.6	21.6	37.3
ResNet SuperGlobal	1024	74.3	83.1	20.5	33.1
DINOv2 SALAD	8192 + 256	**92.2**	**96.4**	**76.0**	**89.2**

Table 9. Results of SuperGlobal (GeM+ + Regional-GeM + Scale-GeM).

SF-XL: Curated from a very large collection of Google Street View images from San Francisco, it contains severe viewpoint and time changes. Its test database contains 2.8M images and provides two different query sets of 1000 and 598 images. Given the size of the test dataset, it serves to evaluate VPR at scale.