Self-Supervised Learning for Place Representation Generalization across Appearance Changes
– Supplementary material –

Mohamed Adel Musallam
mohamed.ali@uni.lu

Vincent Gaudillière
vincent.gaudilliere@uni.lu

Djamila Aouada
djamila.aouada@uni.lu

SnT, University of Luxembourg

In this supplementary material, we provide additional information, analyses and visualizations to support and enhance the understanding of our main paper. The main objective of our paper is to present a novel self-supervised learning framework that effectively generalizes place representations across various appearance changes in diverse environments.

This supplementary material is organized as follows:

Section 1: We go through the technical details of our geometric augmentations for the self-supervised learning method, offering further explanation of the chosen type of transformation, the learning objectives and the optimization techniques employed. This aims at providing the readers with a thorough understanding of the inner working of our approach and its ability to generate robust place representations, as well as the reasoning behind our choices.

Section 2: We provide precisions regarding the reasoning behind appearance augmentations.

Section 3: We share some qualitative visualizations of the learned place representations on the "Oxford RobotCar Seasons v2" dataset, which offers valuable insights into how our framework captures meaningful features and adapts to appearance changes. These visualizations use Grad-CAM heatmaps and example images from different seasons allowing the readers to observe the effectiveness of our method in real-world scenarios.

Section 4: We present findings related to the hypothesis concerning the performance of CLASP-Net when the input query is subjected to rotation.

1. Precisions on Geometric Augmentations

We tested different types of geometric augmentation (see Figure 1). In particular, random affine, random perspective, random rotation, and 90°-rotation have been used in our investigation. However, it has been observed that 90-degree rotation augmentation leads to better performance compared to the other types of augmentations.

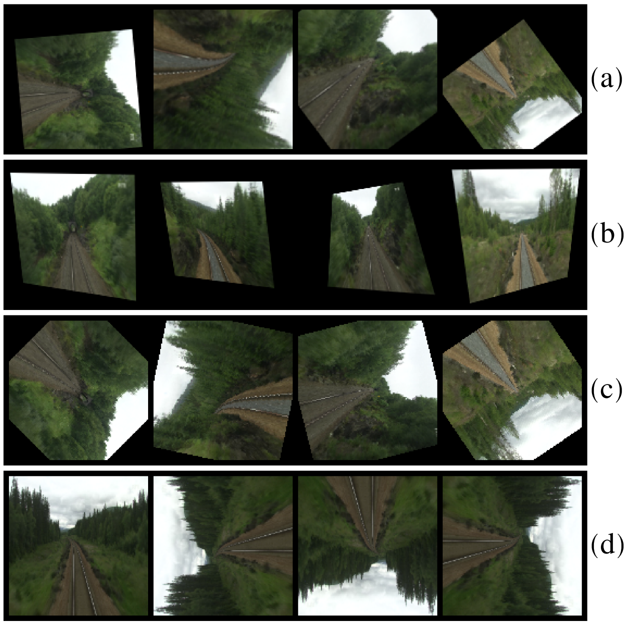

Figure 1. Examples of geometric augmentations: (a) affine transformations, (b) perspective transformations, (c) planar rotations by any angle, (d) planar rotations by 90°, 180°and 270°.

There could be several reasons why the 90-degree rotation augmentation is more effective for visual place recognition.

Random affine and random perspective augmentations can distort the image's semantic content by changing the shape and size of objects, making it harder for the model to recognize the place, primarily since we rely on global features of the input image.

On the other hand, random rotations and 90-degree rotations maintain the spatial structure of the image while providing variations in orientation. However, random continuous rotation may generate artifacts due to the boundary

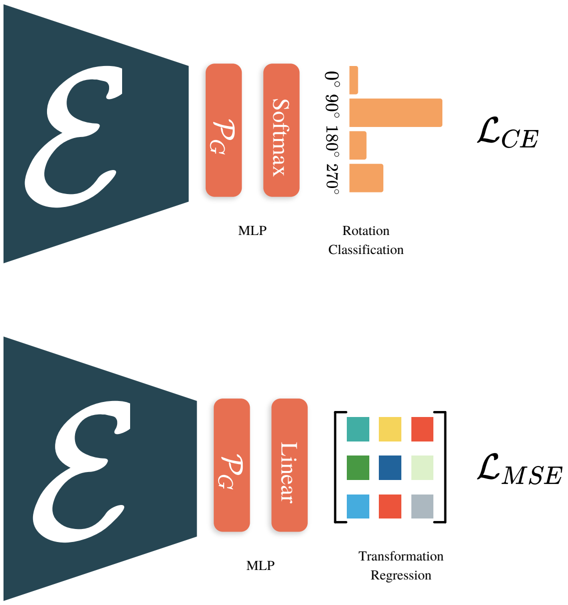

Figure 2. Illustrating the adaptation of the model to the different types of geometric augmentations.

effect. Thus leveraging the group of 90-degree rotations is more helpful for the model to learn spatially meaningful representations, especially since there is no artifacts that the model can learn as shortcuts.

Furthermore, the effectiveness of different augmentations can depend on the dataset's specifics, as mentioned in [1, "Discussion" section]. For our case, natural visual navigation datasets contain a bias in the data towards up (sky) and down (ground), and breaking this natural setting without introducing other bias with the 90-degree rotation augmentation, then encouraging the network to be sensitive to this transformation, can help learning meaningful features.

In conclusion, the 90-degree rotation augmentation is more effective for visual place recognition than other tested geometric augmentations. This could be due to its ability to preserve the image's semantic content and spatial structure, provide variability in orientation, and disturb the natural uprigth setting of the places in the dataset.

The change in the model when learning to regress the parameters of the different types of geometric transformations is illustrated in Figure 2. In particular, the angle of the 90°-rotations is predicted as a classification problem and the training leverages the cross-entropy loss $\mathcal{L}_{CE}$, while the parameters of other transformations are regressed in their matrix form with a Mean Squared Error loss $\mathcal{L}_{MSE}$.

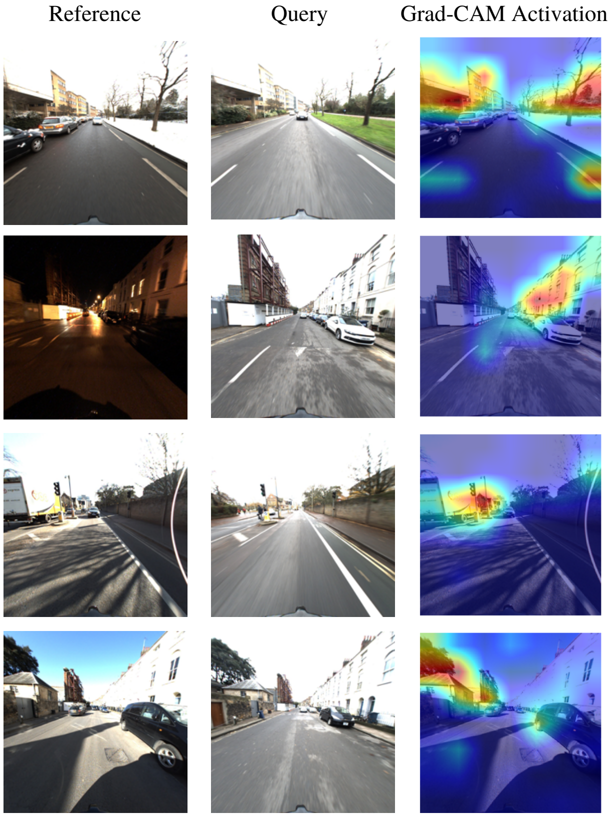

Figure 3. Visual grad-CAM activations of input query images, along with retrieved day images from the Oxford RobotCar Seasons v2 dataset.

2. Precisions on Appearance Augmentations

In the context of visual navigation under appearance changes, pixel augmentation can be especially helpful in learning representations that are robust to changes in lighting, weather, and other factors that can affect the scene's appearance. In addition, by training our model using contrastive learning, the model can explicitly learn to map the same scene under different appearance variations to the same embedding, which can be crucial for successful navigation in real-world scenarios.

3. Qualitative Results on Oxford RobotCar Seasons v2 dataset

Figure 3 shows qualitative results of CLASP-Net on the Oxford RobotCar Seasons v2 dataset, along with Grad-CAM activation maps.

4. The Impact of Rotated Query Images

While for visual place recognition, the orientation of the input image - particularly its natural up and down direction

- is of critical significance, we intentionally break this inherent directionality to help the network learn distinctive locales and attributes of the visual environment.

More precisely, in the rotation prediction task, to recognize a rotated version of a scene, the model needs to identify its different elements (*e.g.* sky, ground, building, objects) and understand the spatial relationships between them [3]. Indeed, the GradCAM visualizations from our paper and supplementary material suggest that CLASP-Net learnt to detect geometric features such as skylines or buildings, which are helpful for both rotation prediction and outdoor place recognition (see [4]).

However, we agree that such sensitivity to rotations may harm the place recognition performance under the same transformations of queries. Addressing such an hypothetical scenario, *i.e.* recognizing the place depicted in a rotated query, therefore requires further adaptation of our original method. We have ran experiments to provide insights towards solving such a problem, whose results are presented below:

(A) **Standard CLASP-Net on Unaltered Queries:** When tested on unrotated queries, the baseline CLASP-Net yielded a performance score of 0.8025. This establishes the primary standard of comparison for the subsequent experiments.

(B) **Standard CLASP-Net on Rotated Queries:** Upon challenging the CLASP-Net with rotated query images, with no retraining there was a marked decrease in performance, registering a score of 0.3729 This significant drop underscores the sensitivity of the model to rotations in its standard configuration.

(C) **CLASP-Net with Rotation Augmentation in the Invariance Module on Unaltered Queries:** By integrating rotation augmentation in the invariance module, the performance on unrotated images slightly diminished to **0.7624**. This suggests that while the model becomes better suited to handle rotations, there might be a slight compromise in its efficacy on standard queries.

(D) **CLASP-Net with Rotation Augmentation in the Invariance Module on Rotated Queries:** Significantly, this configuration exhibited a performance score of **0.7744** on rotated queries. This is a stark contrast to the performance of the standard CLASP-Net on rotated queries, highlighting the efficacy of the rotation augmentation in bolstering the model's resilience against rotations.

(E) **CLASP-Net Excluding the Rotation Prediction Module with Rotation Augmentation in the Invariance Module on Unaltered Queries:** Omitting the equivariance module while retaining the rotation augmentation in the invariance module resulted in a performance of **0.7872** on standard queries.

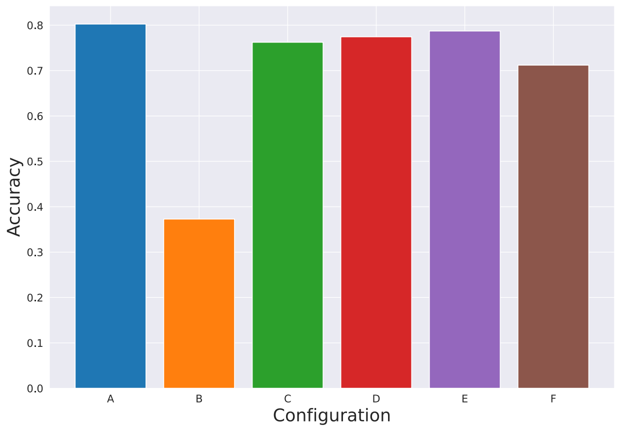

Figure 4. Testing the model with a rotated query in a different configuration on the Nordland Summer/Winter R@10 .

Module	R@10	Equiv. meas. [2]
Pre-trained encoder	28.2%	0.339
Appearance Module	75.8%	0.189
Geometry Module	52.3%	**0.093**
Combined modules (ACM-Net)	**80.2%**	0.139

Table 1. CLASP-Net analysis on Nordland summer/winter.

(F) **CLASP-Net Excluding the Rotation Prediction Module with Rotation Augmentation in the Invariance Module on Rotated Queries:** This configuration produced a score of **0.712** on rotated queries. While this is lower than the performance achieved with the rotation-augmented CLASP-Net on rotated queries, it is considerably better than the standard CLASP-Net's performance on similar inputs.

Our training strategy encourages information about rotations to be retained in the image representation rather than guaranteeing strict equivariance. However, in practice, we observe that the average cosine similarity between representations of rotated views (referred to as *equivariant measure* in [1]) tends to 0 (*i.e.*, 90° angle) when the dedicated module is added (see Table 1). Moreover, the choice of this particular group of geometric transformations is the outcome of experimentation. In particular, it shows that the best performance is achieved with the cyclic group of 90° rotations, compared to the groups of 2D affine transformations, 2D projective transformations, and 2D rotations.

In summary, integrating rotation augmentation within the invariance module substantially improves CLASP-Net's capability to handle rotated queries. However, the omission of the equivariance module has nuanced effects; These insights are instrumental for refining architectures based on the downstream task and the assumption about the input images and enhancing the robustness of the models in the face of varied input conditions.

References

[1] Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljacic. Equivariant self-supervised learning: Encouraging equivariance in representations. In *International Conference on Learning Representations*, 2022. 2, 3

[2] Dangovski et al. Equivariant self-supervised learning: Encouraging equivariance in representations. In *ICLR*, 2022. 3

[3] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 3

[4] Olivier Saurer, Georges Baatz, Kevin Köser, L'ubor Ladický, and Marc Pollefeys. Image based geo-localization in the alps. *International Journal of Computer Vision*, 116(3):213–225, Feb 2016. 3