Self-Supervised Learning for Place Representation Generalization across Appearance Changes

Mohamed Adel Musallam
mohamed.ali@uni.lu

Vincent Gaudillière
vincent.gaudilliere@uni.lu

Djamila Aouada
djamila.aouada@uni.lu

SnT, University of Luxembourg

Abstract

Visual place recognition is a key to unlocking spatial navigation for animals, humans and robots. While state-of-the-art approaches are trained in a supervised manner and therefore hardly capture the information needed for generalizing to unusual conditions, we argue that self-supervised learning may help abstracting the place representation so that it can be foreseen, irrespective of the conditions. More precisely, in this paper, we investigate learning features that are robust to appearance modifications while sensitive to geometric transformations in a self-supervised manner. This dual-purpose training is made possible by combining the two self-supervision main paradigms, i.e. contrastive and predictive learning. Our results on standard benchmarks reveal that jointly learning such appearance-robust and geometry-sensitive image descriptors leads to competitive visual place recognition results across adverse seasonal and illumination conditions, without requiring any human-annotated labels.[1].

1. Introduction

Visual Place Recognition (VPR) is central for localizing - *i.e.* determining a camera's position in a scene [21,40], and has applications from autonomous driving to augmented reality. Typically viewed as an image retrieval task, VPR aims to match a *query* image to images in a *reference* database that depict the same location, even when conditions like viewpoint, obstructions, or weather vary [27]. This makes VPR challenging but vital for dependable real-world vision-based systems.

For this goal, neuroscience research indicates that biological intelligence relies on creating abstract representations of places, known as *cognitive maps* [27], to recognize them under varying conditions [57].
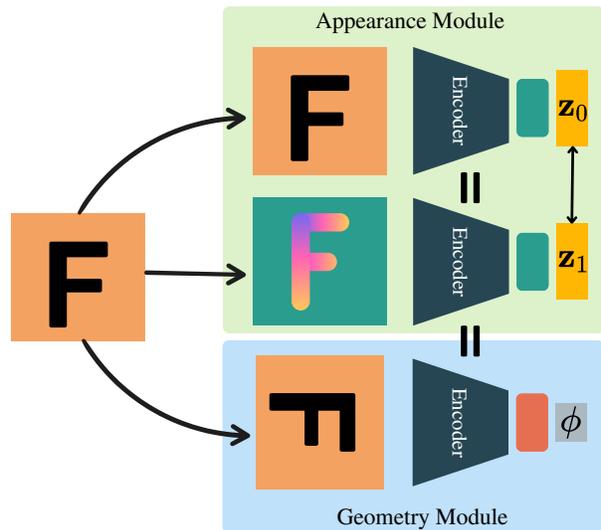
Figure 1. **CLASP-Net Training Strategy:** Three views are generated from an input image. The Appearance Module in green (top) maps the original and appearance-augmented views into close representation vectors $\{z_0, z_1\}$. The Geometry Module in blue (bottom) predicts the transformation ϕ applied between the original and third views.

These maps are essential for generalizing limited knowledge, such as recognizing a place seen only in daylight during nighttime. The aim is to build rich representations reflecting the intrinsic structures that are not required to be re-learned from scratch when non-critical visual information changes [57].

In the context of technological solutions, state-of-the-art VPR methods have focused on achieving invariance to both environmental conditions and viewpoint changes in image representations. The latter are for recognizing places observed under unprecedented angles [11, 26]. However, we argue that such viewpoint invariance may be detrimental in the process of distinguishing between different places.

Moreover, recent works have shown that favouring a more general equivariance in image representations may be more beneficial than seeking *only* invariance [8, 34, 56].

Motivated by this, we introduce **CLASP-Net** :*Contrastive Learning with Appearance Augmentations and Spatial Predictions for Place Recognition*[2], designed to learn discriminative place representations that can generalize to new conditions. We use Self-Supervised Learning (SSL) to address training limitations due to low appearance variability in reference images. Contrastive Learning (CL) is employed to unify representations of the same place by using appearance augmentations. To further exploit the scene's spatial layout and regularize the model, we apply geometric transformations and utilize a Predictive Learning (PL) framework for classification based on these transformations.

Contrary to supervised learning, which tends to learn shortcuts and struggles to generalize from limited labelled data [15], SSL seems closer to human-like learning and does not require manual annotation [25]. Few studies have explored SSL for VPR [12, 48]. We propose to merge the key SSL approaches, CL [6] and PL [22], aiming for image representations that are resilient to appearance shifts while being sensitive to geometric cues. By doing that, we aim to learn features suitable for visual place recognition under appearance changes.

Contributions. Our contributions are two-fold:

(1) A novel approach for Visual Place Recognition under extreme condition changes, CLASP-Net, that leverages both contrastive and predictive self-supervised learning approaches.

(2) An experimental evaluation confirming the competitiveness of CLASP-Net compared to state-of-the-art approaches on standard benchmarks featuring different conditions (day/night, weather, seasons), among which the very challenging Alderley Dataset [31].

Paper organization. The rest of the paper is organized as follows. Relevant work on SSL and VPR is reviewed in Section 2. CLASP-Net is presented in Section 3, while experimental evaluation demonstrating the validity of our approach is reported in Section 4. Section 5 concludes the paper and presents future works.

2. Related Work

2.1. CNN-based Descriptors for Visual Place Recognition

The rapid evolution of deep learning has opened new avenues for overcoming the limitations of traditional, handcrafted descriptors. Following the groundbreaking work by Chen *et al.* [7], there has been a growing emphasis on learning-based descriptors primarily built from Convolutional Neural Networks (CNNs). For instance, Sunderhauf *et al.* [46] and Hou *et al.* [20] found that mid-level features from trained CNN model are more resilient to variations in appearance.

Moreover, a concerted effort has been made to design specialized neural networks for VPR tasks. This has led to the invention of techniques like CALC [29], NetVLAD [1], NetBoW [37] and NetFV [30] that meld the best aspects of both traditional and learning-based descriptors, achieving unprecedented results.

In terms of performance, CNN-based descriptors, particularly those relying on supervised learning, are highly dependent on extensive, high-quality training datasets.

However, it's crucial to acknowledge that supervised learning methods often require laborious data annotation, which can be both time-consuming and costly. Therefore, self-supervised learning presents a compelling alternative to VPR tasks.

2.2. Self-Supervised Learning

Self-supervised methods focus on learning visual features from large sets of unlabeled images, making them valuable for diverse real-world applications such as autonomous driving. These methods usually employ a pretext task with a related objective function for training [22]. The objective function can target either network predictions (predictive learning) or the feature representation space (contrastive learning). This enables SSL to yield image representations that are both sensitive and robust to specific transformations.

Predictive Learning. PL uses pretext tasks to indirectly infuse image representations with inductive biases via network outputs [22]. Tasks range from image colorization [61] and jigsaw puzzles [35] to rotation prediction [16]. These tasks encourage the network to learn rich object representations and their spatial arrangements. For instance, predicting an outdoor scene often involves recognizing sky and trees at the top and roads at the bottom, requiring an understanding of the scene's structure.

Contrastive Learning. CL directly refines image representations using a contrastive loss that considers batch elements' relationships. SimCLR [6], a framework for visual representation through CL, stands out for its simplicity, not needing specialized structures [2] or memory banks [32, 60]. It works by sampling two distinct augmentations, applying each to an image, and then training encoders on a contrastive loss to maximize similarity between the two views and minimize similarity with different images. To address potential training convergence challenges in CL, ScatSim-

---

[2]Clasp: [noun] a device, usually of metal, for fastening together two or more things or parts of the same thing.

CLR [23] also estimates each view's augmentation parameters.

Combining Predictive and Contrastive Learning. CL aims at inducing invariance to some content-preserving transformations while being distinctive to such content changes. On the other side, PL is mostly used to incorporate sensitivity, and ideally equivariance, to given transformations into representations [56]. Some studies have demonstrated the advantage of balancing invariance and equivariance [13, 38, 55]. For example, Winter et al. [59] suggested an AutoEncoder-centric framework to cultivate representations that exhibit both robustness and sensitivity to rotations. Explicitly, an encoder translates a rotated image into a more invariant latent representation, from which a decoder predicts the unrotated original image. Simultaneously, an auxiliary branch pursues equivariance by determining the rotation angle. In a similar vein, Feng *et al.* [10] endeavour to learn features impervious to the rotation of input images by bifurcating the features: one segment is dedicated to rotation prediction (dubbed equivariant features), while another segment, subjected to a contrastive loss, penalizes disparities emerging from various rotations (termed invariant features).

In recent work, Dangovski *et al.* [8] introduced Equivariant Self-Supervised Learning (E-SSL), a more nuanced SSL approach that goes beyond simply seeking invariant representations. E-SSL framework enriches traditional SSL methods by integrating both equivariance and invariance objectives in the pre-training process. The key insight is that some transformations are better captured as equivariant, meaning that the learned features should change predictably based on how the input is transformed. At the same time, other transformations are better captured as invariant, where the feature representation should remain constant despite changes of the input.

Drawing on these insights, our proposed CLASP-Net focuses on achieving appearance invariance through CL while capturing detailed representations of scene components and their spatial layouts through PL. With the latter, the network gains sensitivity to geometric transformations, enhancing its suitability for VPR tasks.

2.3. Self-Supervised Learning for Visual Place Recognition

As highlighted in Section 2.2, SSL is well-suited for VPR because it addresses the issue of unrepresentative training data due to varying test conditions. Despite its promise, few methods exist. For instance, Tang *et al.* [48] have proposed to disentangle appearance-related and place-related features using a generative adversarial network with two discriminators. However, this type of method may suffer from unstable training. SeqMatchNet [12] is a CL-based method that leverages sequences of video frames in the contrastive loss to robustify image representations for VPR.

From a larger perspective, Mithun et al. [33] use sets of related images (i.e., showing the same place under different conditions) to enhance VPR image representations. Thoma et al. [49] suggest loosening geo-tag constraints for weakly-supervised training. Unlike these works, we generate pairs of corresponding images in a self-supervised manner, without labels. Venator et al. [52] employ SSL to create appearance-invariant descriptors for image matching, which could serve as a refinement step in our approach.

3. Proposed CLASP-Net

Our primary objective is to enable the model to learn features that can withstand drastic changes in appearance while remaining effective for VPR. Specifically, we aim to create image representations that capture essential geometric details of the scene's spatial arrangement yet remain unaffected by varying environmental conditions. To accomplish this, we integrate both sensitivity to geometric information and robustness to appearance changes into the image representations using self-supervised learning techniques.

3.1. Problem Formalization

Following the traditional approach [27], we frame the VPR problem as an image retrieval task, where, given a query image $\mathbf{q}$ depicting a place $\mathscr{P}_{\mathbf{q}}$, a representation *a.k.a.* descriptor $\mathbf{z}_{\mathbf{q}}$ of that image is computed. It is then compared to the descriptors $\{\mathbf{z}_i\}_{i=1..N_R}$ of reference images $\{\mathbf{x}_i\}_{i=1..N_R}$, where N_R is the size of the reference database. The comparison is done using a given similarity metric (*e.g.*, cosine similarity). This inference stage is illustrated in Figure 2.

During the training, the model only has access to reference images that we assume unlabelled. Moreover, the environmental conditions under which the query image is acquired are not necessarily similar to the ones featured in the reference database, making the problem very challenging, even sometimes for human eyes.

3.2. Preliminaries: Robustness & Sensitivity

Our approach focuses on extracting image features that are both robust to appearance changes and sensitive to geometric aspects. Mathematically, these properties correspond to the concepts of invariance and equivariance. Formally, let $\mathfrak{G}$ be a generic group of transformations and $\mathfrak{g}$ an element of $\mathfrak{G}$. The actions of $\mathfrak{g}$ on the input and output spaces of a function $\mathcal{F} : \mathbb{I} \to \mathbb{O}$ are denoted by $\phi_{\mathfrak{g}}^{(\mathbb{I})}$ and $\phi_{\mathfrak{g}}^{(\mathbb{O})}$, respectively.

In practice, considering an encoder model $\mathcal{E}$ for extracting features from an image $\mathbf{x}$, we seek robustness to any
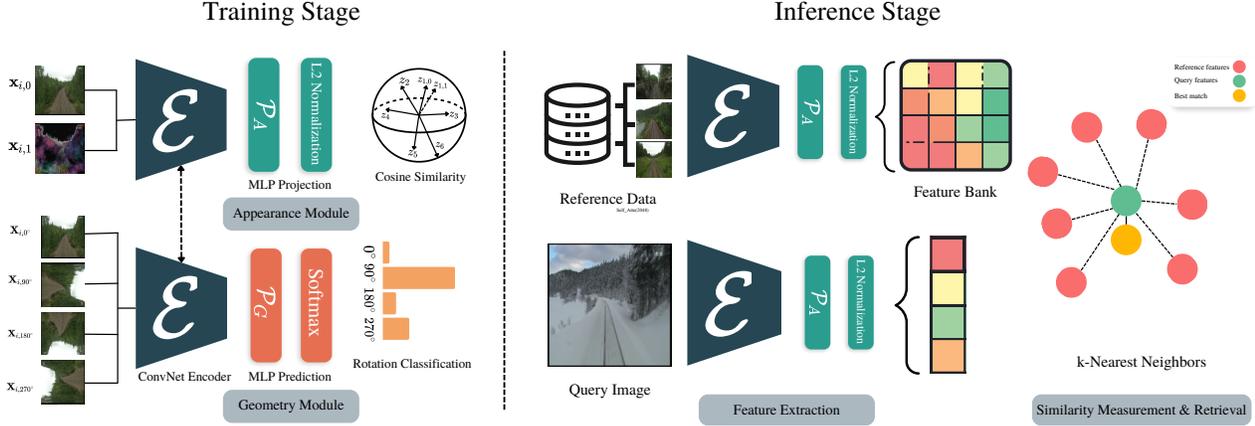
Figure 2. Overview of CLASP-Net. **Training Stage:** from an original image $\mathbf{x}_{i,0}$, augmented versions with a modified appearance $\mathbf{x}_{i,1}$ and different orientations ($\mathbf{x}_{i,0°}$, $\mathbf{x}_{i,90°}$, $\mathbf{x}_{i,180°}$, $\mathbf{x}_{i,270°}$) are generated. Representations of the first two images are brought closer thanks to a contrastive learning framework to achieve appearance robustness. In parallel, original and rotated images are passed through a classification network sharing the same encoder to predict the applied transformation and achieve geometric sensitivity. Note that our method does not rely on any manual annotation. **Inference Stage:** The representations from query and reference images are compared based on similarity measure then the closest k reference images constitute the image retrieval output.

appearance transformation $\mathcal{T}_A$:

$$\forall \mathcal{T}_A, \forall i \in [1; N_R], \quad \mathcal{E}(\mathcal{T}_A \mathbf{x}_i) \approx \mathcal{E}(\mathbf{x}_i), \qquad (1)$$

while, at the same time, sensitivity to a certain group of geometric transformations $\mathfrak{G}_G$:

$$\forall \mathcal{T}_G \in \mathfrak{G}_G, \forall i \in [1; N_R], \quad \mathcal{E}(\mathcal{T}_G \mathbf{x}_i) \approx \mathcal{T}'_G \mathcal{E}(\mathbf{x}_i), \quad (2)$$

where $\mathcal{T}'_G \approx \mathcal{T}_G$. The different possible groups of transformations are investigated in Section 4.

3.3. Model Architecture

Our pipeline exploits both CL for encouraging invariance to appearance changes and PL for encouraging sensitivity to geometric image augmentations. This hybrid approach is consistent with the *E-SSL* framework proposed in [8]. The overall architecture of the proposed CLASP-Net is presented in Figure 2.

At training time, CLASP-Net is composed of two branches sharing the weights of an encoder model $\mathcal{E}$. The first branch, denoted *Appearance Module*, takes as inputs the original image $\mathbf{x}_i$ and an augmented version with modified appearance $\mathcal{T}_A \mathbf{x}_i$, then applies a contrastive learning loss in the representation space to bring the two descriptors closer. The second branch, denoted *Geometry Module*, uses rotated versions of the original image, $\mathcal{T}'_G \mathbf{x}_i = \mathrm{R}(n°)\mathbf{x}_i$, and predicts the angle of the rotation n.

Appearance Module. The first branch, divided into two sub-branches (see Figure 1), This setup is inspired by Sim-CLR [6] and employs a shared encoder $\mathcal{E}$ and MultiLayer Perceptron (MLP) $\mathcal{P}_A$ mapping between the image domain

and the latent representation space where the contrastive loss is applied. Given original images $\mathbf{x}_i$ along with their augmented versions $\mathcal{T}_A \mathbf{x}_i$, the weights of the two networks are learned using a contrastive loss. This loss, formalized in Section 3.4, ensures that the descriptor of each version, *e.g.*, $\mathcal{E}(\mathcal{P}_A(\mathbf{x}_i))$, is similar to the descriptor of its corresponding view, $\mathcal{E}(\mathcal{P}_A(\mathcal{T}_A \mathbf{x}_i))$, while distant from the other descriptors. The intuition behind this module is to force the encoder model $\mathcal{E}$ to learn features agnostic on the conditions (e.g. illumination, weather, season) under which the place was initially observed.

Geometry Module. The second branch incorporates the same shared encoder $\mathcal{E}$ along with a prediction-focused $\mathcal{P}_G$. This setup is designed to classify rotated versions of the original image, denoted $\mathrm{R}(n°)\mathbf{x}$, based on their rotation angle n. Utilizing a standard cross-entropy loss, the module aims to train the encoder $\mathcal{E}$ to learn rich representations of scene layout and spatial arrangement and capture geometry-sensitive features vital for accurate place recognition.

Combined, these two modules work together to disentangle appearance and geometric aspects of input images, enabling robust visual place recognition even when appearance conditions vary. During inference, the architecture used to compute image descriptors consists of the encoder $\mathcal{E}$ followed by the projector network $\mathcal{P}_A$, as shown in Figure 2 (right part).

3.4. Model Loss

Note: For the sake of clarity, we herein introduce more specific notations for denoting images and their augmented/rotated versions.

Figure 3. Examples of augmentations leveraged by CLASP-Net. Top row (a): an original input batch from Oxford RobotCar v2 dataset, (b) pixel-level augmentations for appearance changes, (c) examples of rotations applied to the original image.

We use a combination of contrastive and predictive losses to steer our model toward robustness to appearance changes and sensitivity to geometric variations.

Given a random batch of N reference images $\mathcal{B} = \{\mathbf{x}_{i,0}\}_{i=1..N}$ corresponding to N different places, we apply one random appearance transformation to each image. By so doing, we create N additional images $\{\mathbf{x}_{i,1}\}_{i=1..N}$. These $2N$ images constitute the contrastive batch $\mathcal{B}_C = \{\mathbf{x}_{i,j}\}_{i=1..N, j\in\{0,1\}}$ that is fed into the Appearance Module. Furthermore, we also apply rotations of $0°$, $90°$, $180°$ and $270°$ to each original image. As a result, we create the predictive batch of $4N$ images $\mathcal{B}_P = \{\mathbf{x}_{i,j}\}_{i=1..N, j\in\Theta_4}$, where $\Theta_4 = \{0, 90, 180, 270\}$. $\mathcal{B}_P$ is fed into the Geometry Module.

Contrastive loss. The contrastive batch $\mathcal{B}_C$ contains N *positive* pairs of images $(\mathbf{x}_{i,0}, \mathbf{x}_{i,1})$ depicting the same place, the rest being *negative* pairs corresponding to different places. We use NT-Xent loss [6] that leverages positive samples, and is based on the cosine similarities between the obtained image representations $\mathbf{z}_{\cdot,\cdot} = \mathcal{P}_A(\mathcal{E}(\mathbf{x}_{\cdot,\cdot}))$, expressed as

$$\mathrm{s}(\mathbf{z}_{i,j}, \mathbf{z}_{k,l}) = \frac{\mathbf{z}_{i,j} \cdot \mathbf{z}_{k,l}}{\|\mathbf{z}_{i,j}\|\|\mathbf{z}_{k,l}\|}, \qquad (3)$$

where $\cdot$ is the dot product.
Specifically, the contrastive loss is defined as

$$\mathcal{L}_C = \frac{1}{2N} \sum_{i=1}^{N} \ell_{0\to 1}(i) + \ell_{1\to 0}(i), \qquad (4)$$

where

$$\ell_{a\to b}(i) = -\log \frac{\exp(\mathrm{s}(\mathbf{z}_{i,a}, \mathbf{z}_{i,b})/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{k\neq i} \sum_{j=0}^{1} \exp(\mathrm{s}(\mathbf{z}_{i,a}, \mathbf{z}_{k,j})/\tau)}, \qquad (5)$$

with τ denoting a temperature parameter that controls the strength of penalties on pairs of non-corresponding images [53] and $\mathbb{1}_{k\neq i}$ being equal to 1 if $k \neq i$, and 0 otherwise.

The contrastive loss aims at making representations of the same place under different conditions similar to each other, while forcing representations of different places to be different.

Predictive loss. The predictive batch $\mathcal{B}_\mathcal{P}$ contains four rotated views of each place. The task of this branch is to predict the rotation angle for each of the $4N$ pictures. We frame this as a classification problem with 4 classes corresponding to $0°$, $90°$, $180°$ and $270°$ rotation angles. The predictive loss is therefore the standard cross-entropy loss:

$$\mathcal{L}_P = -\sum_{i=1}^{N} \sum_{j\in\Theta_4} \mathrm{c}(\mathbf{x}_{i,j}) \cdot \log(\widetilde{\mathbf{z}}_{i,j}), \qquad (6)$$

where $\widetilde{\mathbf{z}}_{i,j} = \mathrm{Softmax}(\mathcal{P}_G(\mathcal{E}(\mathbf{x}_{i,j}))) \in \mathbb{R}^4$ is the prediction, $\log()$ the element-wise natural logarithm, $\cdot$ the dot product and $\mathrm{c}(\mathbf{x}_{i,j}) \in \mathbb{R}^4$ the groundtruth with elements equal to 0 except the nth element equal to 1 if the true rotation is $(n-1) \times 90°$.

Overall loss. The final loss is the combination of the contrastive loss for appearance robustness and predictive loss for geometry sensitivity:

$$\mathcal{L} = \mathcal{L}_C + \lambda.\mathcal{L}_P, \qquad (7)$$

Data Augmentation Type	Probability
Planckian Jitter	0.8
Color Jiggle	0.5
Plasma Brightness	0.5
Plasma Contrast	0.3
Gray scale	0.3
Box Blur	0.5
Channel Shuffle	0.5
Motion Blur	0.3
Solarize	0.5

Table 1. List of data augmentations applied to the images on-the-fly during training. We also set a probability for each one of them.

where λ is a weighting factor to balance the two terms.

4. Experimental Evaluation

4.1. Datasets

The Nordland dataset [45]: records a 728 km long train journey connecting the cities of Trondheim and Bodø in Norway. It contains four long traversals, once per season, with diverse visual conditions. The dataset has 35768 images per season with one-to-one correspondences between them. We follow the dataset partition proposed by Olid *et al.* [36] with test set made of 3450 photos from each season.

The Alderley dataset [31]: records an 8 km travel along the suburb of Alderley in Brisbane, Australia. The dataset contains two sequences: the first one was recorded during a clear morning, while the second one was collected on a stormy night with low visibility, which makes it a very challenging benchmark. The dataset contains 14607 images for each sequence and each place have 2 images. We train our approach on the day sequence and test on the night sequence.

The Oxford RobotCar Seasons v2 dataset [50]: is based on the RobotCar dataset [28], which depicts the city of Oxford, UK. It contains images acquired from three cameras mounted on a car. There are 10 sequences corresponding to 10 different traversals carried out under very different weather and seasonal conditions. The rear camera images of the *overcast-reference* traversal (6954 images) are used as a basis for reference training images, to which we add 1906 rear camera images from other traversals following the *v2* train/test split. These additional images cover different environmental conditions but only a subset of places (not full traversals). The test set contains 1872 images from all traversals except *overcast-reference*, without overlap with training images.

4.2. Evaluation

The evaluation on both Nordland and Alderley datasets uses the recall R@N measure, which consists in the pro-

Method	Nordland Summer/Winter		
	R@1	R@5	R@10
NetVLAD [1]	7.7	13.7	17.7
SFRS [14]	18.8	32.8	39.8
SuperGlue [43]	29.1	33.5	34.3
DELG [4]	51.3	66.8	69.8
Patch-NetVLAD [17]	46.4	58.0	60.4
TransVPR [54]	**58.8**	**75.0**	*78.7*
CLASP-Net (Ours)	*53.0*	*73.8*	**80.2**

Table 2. Quantitative results on Nordland dataset. Best results are in **bold**. Second best results are in *italic*.

Method	Alderley Day/Night
NetVLAD [1]	3.35
CIM [9]	7.82
Patch-NetVLAD [17]	7.99
Seqslam [31]	9.90
Retrained NetVLAD [47]	15.8
AFD [47]	21.0
CLASP-Net (Ours)	**25.2**

Table 3. Quantitative results on Alderley dataset. Best result is in **bold**.

portion of successfully localized query images when considering the first N retrievals. If at least one of the top N reference images is within a tolerance window around the query's ground truth correspondence, the query image is deemed succesfully localized. The tolerance window is set to two frames distant from the query before and after, so that the window contains 5 pictures. Following the common approach for NordLand [3, 17, 18], images of the winter sequence are used as queries, while the summer sequence is used as reference.

For RobotCar-Seasons v2, we follow the Patch-NetVLAD [17] approach and utilize the 6-DoF pose of the best-matched reference picture as prediction of the query's pose. Since we don't compute any pose, our image retrieval method is not comparable with pose estimation methods such as MegLOC [39].

4.3. Implementation details

Encoder model $\mathcal{E}$. We use ResNet50 [19] as the backbone, with pre-training on ImageNet using the Timm library [58]. The last classification layer is discarded so that the model is only used for the feature extraction.

Rotation predictor $\mathcal{P}_G$. We use a simple 1-layer perceptron with layer normalization and ReLU activation.

Projector $\mathcal{P}_A$ We use a simple 1-layer perceptron with batch normalizations and ReLu activation. The dimension
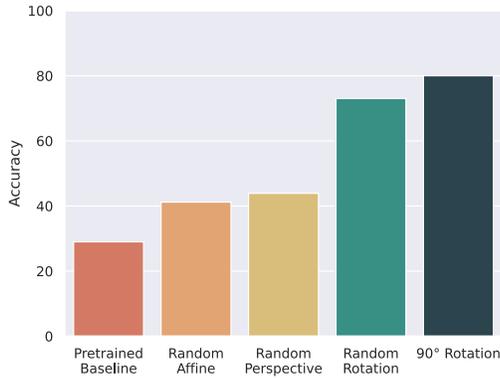
Figure 4. R@10 on Nordland Summer/Winter dataset with Geometry Modules relying on different groups of transformations.

of the output (*i.e.*, image descriptor) is 1024.

Appearance Augmentations. Following domain generalization approaches, our model leverages numerous pixel-level data augmentations to trigger appearance invariance bias in the model. The list of pixel-level augmentations for appearance modification is provided in Table 1, while examples of such augmentations are provided in Figure 3. The chosen set of variations empirically achieved good performance whereas other tested combinations were less favourable. We use the Kornia [42] library for self-supervised data augmentation.

Geometric Augmentations. Our training strategy encourages information about rotations to be retained in the image representation rather than guaranteeing strict equivariance. Moreover, the choice of this particular group of geometric transformations is the outcome of experimentations whose results are presented in Figure 4. Empirically, we found 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.

Model training. The model is trained for 1000 epochs using Adam optimizer [24] and a batch size of 64. Although contrastive learning usually requires larger batch size [5], using Adam optimizer allowed us to obtain good results with a smaller batch size. A learning rate of 0.003 had the best performance with this optimizer. The temperature parameter τ is set to 0.01 and the loss factor λ is set to 1 in our experiments.

Inference. Prior to the inference stage, we pass the set of reference images to the Appearance Invariant Module of the

trained model: $\mathcal{E} \rightarrow \mathcal{P}_A \rightarrow L2-\text{normalization}$ and thus build a reference descriptor bank. A k-Nearest Neighbor search based on cosine similarity to find the closest references to the query image.

4.4. Results

Tables 2, 3 and 4 show the results of CLASP-Net along with other approaches on the three previously described datasets: partitioned Nordland, Alderley Day/Night and RobotCar-Seasons datasets.

The results demonstrate that our method outperforms, by a large margin, standard baselines such as NetVLAD [1] and even local feature-based methods such as Super-Glue [43]. It outperforms Patch-NetVLAD [17] on Nordland dataset (Table 2) and competes with it on Robotcar Seasons v2 (Table 4), despite the fact that Patch-NetVLAD leverages multi-scale descriptors whereas we rely on a single global descriptor. Only the transformer-based architecture TransVPR [54] presents a higher performance as compared to CLASP-Net. We note, however, that our model is based on simple ConvNet and MLP elements that can be upgraded to improve the performance. Finally, it is worth noting that we achieve state-of-the-art results on the very challenging Alderley dataset (Table 3).

Qualitative results are presented in Figure 5 (Nordland dataset) and 6 (Alderley). More qualitative results are included in supplementary materials. One can see examples of queries and best retrieved images, along with Grad-CAM [44] activations. These visualizations demonstrate that CLASP-Net, even if trained without any labels, was able to learn features meaningful for outdoor localization tasks such as skylines for instance.

We focused our study on learning global visual representations that are robust to appearance changes and suitable for VPR. Our results demonstrate that it is possible to learn a model relying on constrastive self-supervision for robustness to appearance changes while being able to perceive the geometric structure of the input image by enforcing geometric prediction.

4.5. Discussion on Potential Limitations

Global image descriptors typically offer greater robustness to environmental conditions at the expense of being less tolerant to viewpoint changes compared to local descriptors [27]. Our approach aims to further enhance the robustness to environmental conditions, allowing it to handle extreme scenarios as seen in the Nordland or Alderley datasets effectively. However, it is important to acknowledge that our method may encounter limitations when dealing with datasets that feature significant viewpoint variations between reference and query images for the same location, as our slightly weaker performance on the Oxford RobotCar dataset suggests.

	day conditions							night conditions	
	dawn	dusk	OC-summer	OC-winter	rain	snow	sun	night	night-rain
m	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0	.25 / .50 / 5.0
deg	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10	2 / 5 / 10
AP-GEM [41]	1.4 / 14.2 / 65.9	9.6 / 29.4 / 82.9	2.4 / 19.1 / 80.5	3.6 / 20.3 / 78.1	4.4 / 21.5 / 86.0	4.5 / 15.8 / 75.9	1.8 / 7.5 / 58.2	0.0 / 0.2 / 6.8	0.1 / 1.2 / 15.8
DenseVLAD [51]	4.5 / 24.3 / 79.6	12.5 / 38.9 / 89.1	3.8 / 27.4 / 90.8	4.1 / 27.1 / 85.6	5.4 / 29.0 / 91.4	6.7 / 25.5 / 85.1	3.2 / 11.0 / 67.1	**1.4** / 2.7 / 23.2	0.6 / 5.2 / 29.8
NetVLAD [1]	2.2 / 16.8 / 73.3	11.4 / 31.0 / 85.9	3.2 / 21.5 / 90.9	4.1 / 22.6 / 84.0	4.2 / 22.2 / 89.4	5.2 / 20.1 / 80.8	2.4 / 10.4 / 70.3	0.2 / 1.2 / 9.1	0.3 / 0.9 / 8.8
DELG global [4]	1.6 / 10.9 / 66.4	8.9 / 23.9 / 81.3	2.1 / 16.5 / 77.6	3.5 / 18.5 / 73.6	3.9 / 20.5 / 87.9	3.6 / 13.5 / 73.5	1.0 / 6.4 / 59.6	0.2 / 0.7 / 7.6	0.1 / 1.6 / 13.8
DELG local [4]	1.7 / 10.4 / 78.3	2.5 / 7.3 / 76.8	1.1 / 8.9 / 84.2	1.2 / 9.1 / 83.2	1.2 / 4.5 / 76.8	3.5 / 10.9 / 80.8	3.3 / 12.6 / 85.2	**1.4** / 7.6 / **38.6**	**2.4** / *11.9* / **53.0**
SuperGlue [43]	4.3 / 24.6 / 84.8	12.7 / 40.3 / 88.6	5.0 / 31.5 / *95.0*	*4.5* / 30.2 / 88.6	5.9 / 30.1 / 91.8	7.0 / 25.4 / 87.2	3.3 / 17.1 / 83.9	0.5 / 2.2 / 27.9	0.9 / 5.4 / *31.8*
Patch-NetVLAD [17]	4.8 / **72.5** / 86.2	**13.5** / **72.0** / 89.5	5.3 / **80.9** / 94.5	**6.3** / **71.3** / 89.8	5.9 / **79.3** / 92.1	7.8 / **75.9** / 87.9	4.8 / **67.3** / 83.4	0.5 / **12.4** / 24.9	*1.0* / **19.0** / 30.8
TransVPR [54]	**18.5** / 52.0 / **95.6**	10.7 / 44.7 / **100.0**	**12.3** / *45.5* / **99.1**	1.2 / *36.6* / **99.4**	**15.1** / 50.7 / **99.5**	**14.0** / *42.8* / **99.1**	**13.4** / *34.4* / **91.1**	*0.9* / 4.9 / *30.5*	0.0 / 1.0 / 10.3
CLASP-Net (Ours)	8.4 / 26.9 / *88.1*	5.1 / 25.9 / 89.8	7.1 / 32.7 / 84.4	0.6 / 22.6 / *91.5*	*12.7* / 42.9 / 93.7	8.8 / 31.2 / 90.2	8.9 / 22.3 / 76.8	0.0 / 2.3 / 14.0	0.0 / 3.0 / 14.8

Table 4. Quantitative results on RobotCar Seasons v2 dataset. Best results are in **bold**. Second best results are in *italic*.
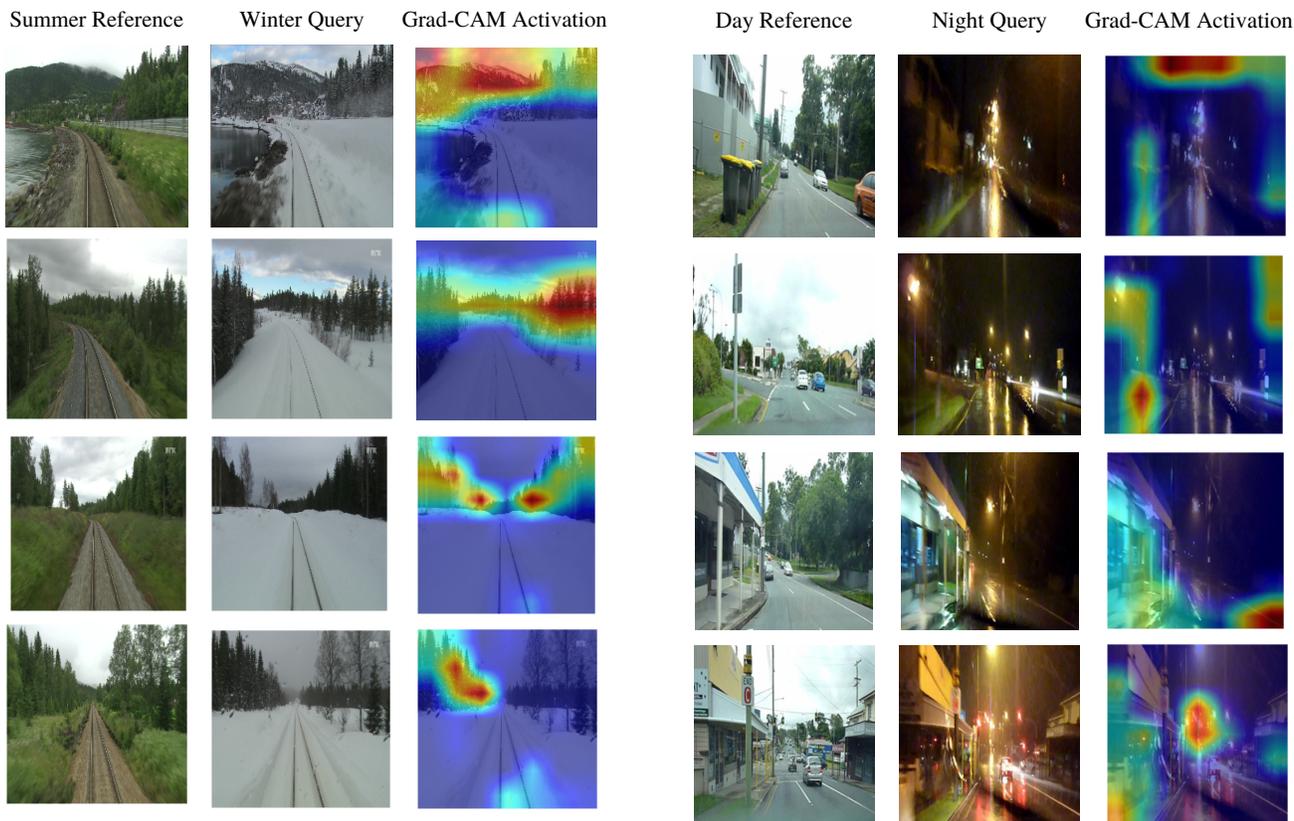

Figure 5. Visual Grad-CAM activation of input query winter image, along with retrieved summer image from the Nordland dataset.

Figure 6. Visual Grad-CAM activation of input query night image, along with retrieved day image from the Alderley dataset.

5. Conclusions

In this paper, we presented CLASP-Net, a novel self-supervised approach designed for visual place recognition under challenging appearance variations. A significant advantage of our method is its independence from human supervision. CLASP-Net is trained to learn features that are both robust to appearance changes and sensitive to geometric nuances, serving as abstract place representations useful for visual place recognition tasks. Our extensive experimental evaluations substantiate the effectiveness and effi-

ciency of the proposed approach. As a direction for future research, we aim to extend our model's capabilities by exploring sensitivity to 3D geometric transformations through view synthesis techniques.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2, 6, 7, 8

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[3] Luis G Camara and Libor Přeučil. Visual place recognition by spatial matching of high-level cnn features. *Robotics and Autonomous Systems*, 133:103625, 2020. 6

[4] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *European Conference on Computer Vision*, pages 726–743. Springer, 2020. 6, 8

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 7

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. 2, 4, 5

[7] Zetao Chen, Obadiah Lam, Adam Jacobson, and Michael Milford. Convolutional neural network-based place recognition. *arXiv preprint arXiv:1411.1509*, 2014. 2

[8] 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, 4

[9] Jose M Facil, Daniel Olid, Luis Montesano, and Javier Civera. Condition-invariant multi-view place recognition. *arXiv preprint arXiv:1902.09516*, 2019. 6

[10] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[11] Sourav Garg, Tobias Fischer, and Michael Milford. Where is your place, visual place recognition? In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4416–4425. ijcai.org, 2021. 1

[12] Sourav Garg, Madhu Babu Vankadari, and Michael Milford. Seqmatchnet: Contrastive learning with sequence matching for place recognition & relocalization. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Conference on Robot Learning, 8-11 November 2021, London, UK*, volume 164 of *Proceedings of Machine Learning Research*, pages 429–443. PMLR, 2021. 2, 3

[13] Quentin Garrido, Laurent Najman, and Yann Lecun. Self-supervised learning of split invariant equivariant representations. *arXiv preprint arXiv:2302.10283*, 2023. 3

[14] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *European conference on computer vision*, pages 369–386. Springer, 2020. 6

[15] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020. 2

[16] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 2

[17] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14141–14152, June 2021. 6, 7, 8

[18] Stephen Hausler and Michael Milford. Hierarchical multiprocess fusion for visual place recognition. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3327–3333. IEEE, 2020. 6

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016. 6

[20] Yi Hou, Hong Zhang, and Shilin Zhou. Convolutional neural network-based image representation for visual loop closure detection. In *2015 IEEE international conference on information and automation*, pages 2238–2245. IEEE, 2015. 2

[21] Martin Humenberger, Yohann Cabon, Noé Pion, Philippe Weinzaepfel, Donghwan Lee, Nicolas Guérin, Torsten Sattler, and Gabriela Csurka. Investigating the role of image retrieval for visual localization. *Int. J. Comput. Vis.*, 130(7):1811–1836, 2022. 1

[22] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4037–4058, 2021. 2

[23] Vitaliy Kinakh, Olga Taran, and Svyatoslav Voloshynovskiy. Scatsimclr: Self-supervised contrastive learning with pretext task regularization for small-scale datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1098–1106, October 2021. 3

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[25] Yann LeCun and Ishan Misra. Self-supervised learning: The dark matter of intelligence. https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/, March 2021. Consulted: October, 2022. 2

[26] Stephanie Lowry and Henrik Andreasson. Lightweight, viewpoint-invariant visual place recognition in changing environments. *IEEE Robotics and Automation Letters*, 3(2):957–964, 2018. 1

[27] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2015. 1, 3, 7

[28] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. 6

[29] Nate Merrill and Guoquan Huang. Lightweight unsupervised deep loop closure. In *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018. 2

[30] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017. 2

[31] Michael J Milford and Gordon F Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *2012 IEEE international conference on robotics and automation*, pages 1643–1649. IEEE, 2012. 2, 6

[32] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[33] Niluthpol Chowdhury Mithun, Cody Simons, Robert Casey, Stefan Hilligardt, and Amit K. Roy-Chowdhury. Learning long-term invariant features for vision-based localization. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pages 2038–2047. IEEE Computer Society, 2018. 3

[34] Mohamed Adel Musallam, Vincent Gaudillière, Miguel Ortiz del Castillo, Kassem Al Ismaeil, and Djamila Aouada. Leveraging equivariant features for absolute pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6876–6886, 2022. 2

[35] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pages 69–84. Springer, 2016. 2

[36] Daniel Olid, José M. Fácil, and Javier Civera. Single-view place recognition under seasonal changes. In *PPNIV Workshop at IROS 2018*, 2018. 6

[37] Eng-Jon Ong, Syed Sameed Husain, Mikel Bober-Irizar, and Miroslaw Bober. Deep architectures and ensembles for semantic video classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(12):3568–3582, 2018. 2

[38] Mandela Patrick, Yuki M. Asano, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9577–9587, October 2021. 3

[39] Shuxue Peng, Zihang He, Haotian Zhang, Ran Yan, Chuting Wang, Qingtian Zhu, and Xiao Liu. Megloc: A robust and accurate visual localization pipeline. *CoRR*, abs/2111.13063, 2021. 6

[40] Noé Pion, Martin Humenberger, Gabriela Csurka, Yohann Cabon, and Torsten Sattler. Benchmarking image retrieval for visual localization. In Vitomir Struc and Francisco Gómez Fernández, editors, *8th International Conference on 3D Vision, 3DV 2020, Virtual Event, Japan, November 25-28, 2020*, pages 483–494. IEEE, 2020. 1

[41] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5107–5116, 2019. 8

[42] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Winter Conference on Applications of Computer Vision*, 2020. 7

[43] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 6, 7, 8

[44] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 7

[45] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. In *Proc. of workshop on long-term autonomy, IEEE international conference on robotics and automation (ICRA)*, page 2013, 2013. 6

[46] Niko Sünderhauf, Sareh Shirazi, Feras Dayoub, Ben Upcroft, and Michael Milford. On the performance of convnet features for place recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4297–4304. IEEE, 2015. 2

[47] Li Tang, Yue Wang, Qianhui Luo, Xiaqing Ding, and Rong Xiong. Adversarial feature disentanglement for place recognition across changing appearance. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1301–1307. IEEE, 2020. 6

[48] Li Tang, Yue Wang, Qimeng Tan, and Rong Xiong. Explicit feature disentanglement for visual place recognition across appearance changes. *International Journal of Advanced Robotic Systems*, 18(6):17298814211037497, 2021. 2, 3

[49] Janine Thoma, Danda Pani Paudel, and Luc V Gool. Soft contrastive learning for visual localization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11119–11130. Curran Associates, Inc., 2020. 3

[50] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, et al. Long-term visual localization revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 6

[51] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view

synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2015. 8

[52] Moritz Venator, Yassine El Himer, Selcuk Aklanoglu, Erich Bruns, and Andreas K. Maier. Self-supervised learning of domain-invariant local features for robust visual localization under challenging conditions. *IEEE Robotics Autom. Lett.*, 6(2):2753–2760, 2021. 3

[53] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, June 2021. 5

[54] Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. Transvpr: Transformer-based place recognition with multi-level attention aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13648–13657, June 2022. 6, 7, 8

[55] Tan Wang, Qianru Sun, Sugiri Pranata, Karlekar Jayashree, and Hanwang Zhang. Equivariance and invariance inductive bias for learning from insufficient data. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings*. Springer, 2022. 3

[56] Yifei Wang, Zhengyang Geng, Feng Jiang, Chuming Li, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Residual relaxation for multi-view representation learning. *Advances in Neural Information Processing Systems*, 34:12104–12115, 2021. 2, 3

[57] James C. R. Whittington, David McCaffary, Jacob J. W. Bakermans, and Timothy E. J. Behrens. How to build a cognitive map. *Nature Neuroscience*, 25(10):1257–1272, Oct 2022. 1

[58] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 6

[59] Robin Winter, Marco Bertolini, Tuan Le, Frank Noé, and Djork-Arné Clevert. Unsupervised learning of group invariant and equivariant representations. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, November 28-December 9, 2022, hybrid*, 2022. 3

[60] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[61] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pages 649–666. Springer, 2016. 2