




DaCo: domain-agnostic contrastive learning for visual place recognition

Hao Ren¹ · Ziqiang Zheng² · Yang Wu³ · Hong Lu¹ 

Accepted: 10 April 2023 / Published online: 14 June 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Visual place recognition is a core component of visual information analysis, which serves for the position and orientation perception of autonomous driving and robotics. The current place recognition methods usually rely on image retrieval techniques to identify the visual similarity between query and gallery images. However, state-of-the-art image retrieval methods are often based on extensive labels, such as matched pairs (e.g., the image correspondences). Besides, image retrieval methods heavily suffer from environmental condition changes (i.e., a large range of illumination and weather changes). To alleviate the annotation cost, we introduce contrastive learning to perform feature extraction and feature similarity measurement in a self-supervised manner. Considering the heavy data augmentations of the existing contrastive learning approaches cannot effectively simulate domain disparities, we design the generative adversarial model to promote the extraction of domain-agnostic features. To tightly integrate the domain-agnostic representations and self-supervision, we design a self-generated soft constraint to achieve domain-agnostic contrastive learning (termed “DaCo”). Extensive experiments and analysis on cross-illumination and cross-weather settings are conducted on three challenging datasets. The proposed “DaCo” outperforms current contrastive learning based image retrieval methods by a large margin.

Keywords Image retrieval · Contrastive learning · Generative adversarial networks · Domain adaptation · Visual place recognition

1 Introduction

Visual place recognition [1, 2] plays a crucial role in the visual information analysis research field, which is widely adopted

in autonomous driving [3] and robotic perception [4]. There are various categories of place recognition methods: Map-based [5], SLAM-based [6], and Image-based [7]. Though the Map-based and the SLAM-based methods can bring a higher place recognition precision with 3D correspondences, they usually require the GPS-tagged annotations. While the later Image-based methods [8, 9] do not require the GPS-tagged annotations, they often rely on the query-gallery matched pairs to accomplish place recognition by directly comparing the query image against the collected image database. Considering it is expensive and time-consuming to obtain these matched image pairs, an invaluable problem arises: how can we perform image-based place recognition without the matched image pairs? Recently, contrastive learning seemingly provides a feasible solution to address this problem.

Due to its inherent superiority of being annotation-free [10], contrastive learning has received a lot of attention. The key concept of contrastive learning is to adopt self-generated supervision to obtain efficient feature representations for downstream vision tasks. Through the contrastive

Hao Ren and Ziqiang Zheng are authors contributed equally to this work.

✉ Hong Lu
honglu@fudan.edu.cn
Hao Ren
hren17@fudan.edu.cn
Ziqiang Zheng
zhengziqiang1@gmail.com
Yang Wu
wu.yang.8c@kyoto-u.ac.jp

¹ Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China

² UISEE Technology Co., Ltd., Shanghai, China

³ Kyoto University, Kyoto, Japan

learning, recent methods such as MoCo [11], SimCLR [12] and SimCLR-v2 [13] have achieved competitive recognition performance compared with the state-of-the-art supervised methods [14, 15]. However, there is still a huge challenge to introduce contrastive learning methods to perform image retrieval tasks in the real world. The visual variability caused by different illumination and various weather shifting significantly influences the accuracy of visual place recognition, which significantly limits its real applications. Thus, directly performing contrastive learning method SimCLR[12] for cross-domain (cross-illumination and cross-weather) image retrieval leads to poor results as shown in Fig. 1.

The essential problem of contrastive learning methods is that they fail to simulate the domain variances through simple data augmentation techniques, such as random resizing or color jittering. Many objects (e.g., the trees and street lamps) look differently across different illumination and weather conditions. For example, the foggy image in Fig. 1 (b) is too fuzzy to extract robust feature representations through only self-generated supervision. Thus, to alleviate this problem, we should promote the ability to extract image representations as robust as possible to visual condition changes. In other words, the learned feature representations are supposed to be domain-agnostic [16]. To achieve this, we introduce the generative model [17] to enhance the ability to extract the domain-agnostic representations.

In detail, we propose a novel, simple and effective **Domain-agnostic Contrastive learning model** (termed **DaCo**) to tightly integrate the contrastive learning and the unpaired domain translation. To perform the feature disentanglement among different domains, we adopt the unpaired image-to-image translation method [18, 19] to eliminate the domain-specific information and preserve domain-agnostic features. Different from the pipeline architecture (“GAN + Recognition”) proposed in [19, 20], we design a self-generated soft constraint to couple the domain-agnostic feature representations and contrastive learning, which leads to a huge performance gain. Our work is also different from the current

i-Mix [21] and DACL [22], which both mainly utilize the Mixup [23] as a data augmentation tool to enhance the performance in different domains.

Unlike the existing visual place recognition algorithms (i.e., EGT [24], Patch-NetVLAD [25]) to obtain the precise camera poses (the translation and rotation matrix), the global-view representations extracted using contrastive learning approaches cannot be served to obtain the detailed geometry constraint (e.g., the patch or pixel correspondences). Existing self-supervised methods are ineffective at obtaining precise patch or pixel correspondences. Thus, in this paper, we are mainly discussing how to obtain image-level correspondences rather than camera poses under a large range of illumination and weather changes. Images with different illumination and weather appearances are regarded from different domains. To our knowledge, the proposal is the first to combine unpaired image translation and unsupervised contrastive learning to perform visual place recognition. To sum up, our main contributions can be summarized as follows:

- The unpaired domain translation is introduced to reduce the domain gap between different domains and the contrastive learning is aimed to extract domain-agnostic features between the images of different domains. They are tightly integrated to produce better retrieval-based visual place recognition performance across different illumination and weather settings.
- A self-generated soft constraint is designed to better align the translated domain and original real domain, which can alleviate the noise and uncertainty caused by the unpaired domain translation.
- Comprehensive experiments including the ones under cross-illumination and cross-weather retrieval-based visual place recognition settings have been conducted. The experimental results have demonstrated the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section 2 reviews some related works. We introduce the details of the

Fig. 1 The examples of cross-domain image retrieval performed by SimCLR [12]: (a) Night → Day and (b) Foggy → Clear. Images with different illumination and weather appearances are regarded from different domains. SimCLR cannot obtain precise image retrieval results due to the illumination and weather changes. Best viewed in color



proposed method in Section 3. The results of experimental evaluation are reported in Section 4, followed by the conclusion in Section 5.

2 Related work

2.1 Retrieval-based place recognition

Retrieval-based place recognition task [24, 26] has long been studied since the era of hand-crafted image descriptors, e.g., SIFT [27], BoW [28] and VLAD [29]. Thanks to the development of deep learning, NetVLAD [30] successfully transformed dense CNN features for place recognition by proposing a learnable VLAD layer to effectively aggregate local descriptors with learnable semantic centers. Adopting NetVLAD as backbone, later works further looked into multi-scale contextual information [25] or effective metric learning [31] to achieve better performance. Some works introduced motion information [2] or geographical verification [24] to help feature learning.

The above-mentioned methods require image-level annotation to perform supervised learning and mostly focus on obtaining the precise camera poses. Compared with these methods, the proposed DaCo is easier to implement because it requires no image-level annotations and focuses on obtaining the image-level correspondences.

2.2 Contrastive learning

The contrastive learning methods [11–13, 32, 33] have achieved a tremendous amount of success in recent years and obtained competitive recognition performance without any human annotation. The key innovation of contrastive learning is to design self-generated supervision and obtain learned feature representations through pretext tasks, such as NPID [32], CMC [33], MoCo [11], SwAV [34], BYOL [35], NNCLR [36] and SimSiam [37].

The representative SimCLR [12] applies different data augmentation techniques to an input image, the transformed outputs are regarded as the same instance (positive samples) while other images are negative samples. By minimizing the global representation distance between different transformed samples and enlarging the distance with the negative pairs, SimCLR [12] could learn to extract the global content information of one image without any label annotation. However, the self-supervised methods fail to capture the domain appearance variance and extract the domain-agnostic features. To conduct the disentanglement of the domain-agnostic and domain-specific feature representations, unpaired domain translation methods could be introduced.

2.3 Unpaired domain translation

Unpaired domain translation [38] between two domains aims to bridge the knowledge between the source domain and the target domain based on unpaired data. The Cycle-GAN methods [18, 19] become popular, which adopt the cycle-consistency loss to preserve the content information during translation procedures.

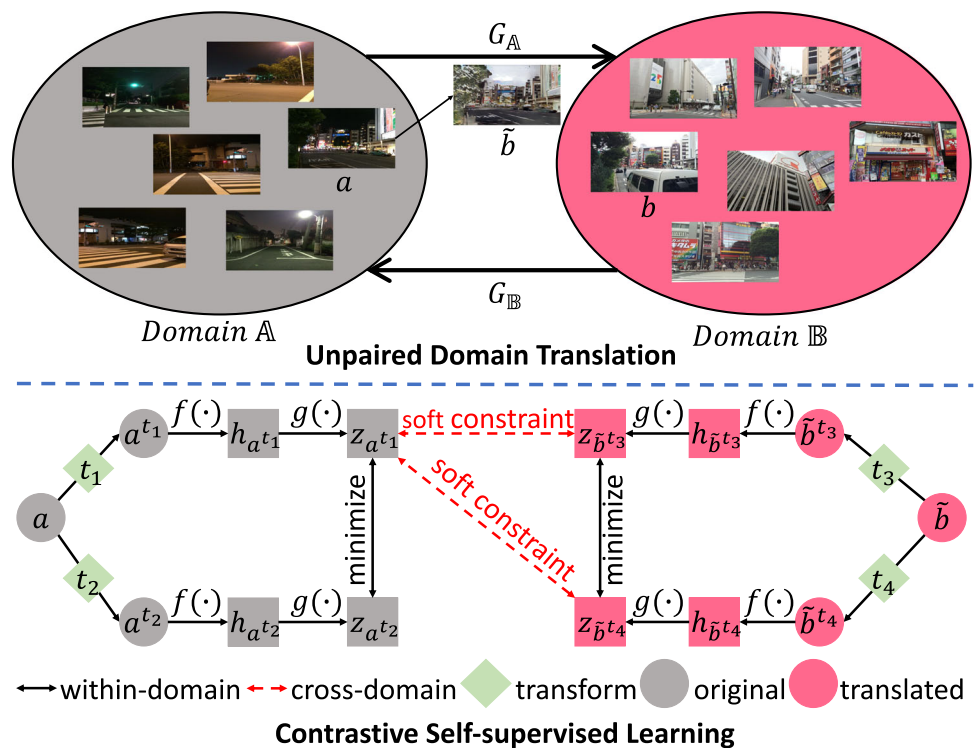
By modeling the mapping function between the source and target domains, [19, 20, 39] aim to capture the domain-agnostic features to boost the performance for complex tasks: place recognition [20], semantic segmentation [19] and object detection [39]. The very similar ToDayGAN [20] combines the unsupervised adversarial domain translation to conduct the night-to-day image translation to boost the supervised image retrieval performance. [19, 20, 39] provide a prototype to combine the unpaired domain translation and recognition tasks through a pipeline architecture. Inspired by these approaches, we design one domain-agnostic contrastive learning method to perform cross-domain image retrieval without matched image pairs.

3 Methods

3.1 Preliminary

In this paper, we aim to find the image of the same place in another domain (i.e., daytime) through the given image in current domain (i.e., nighttime). We use the unpaired domain translation and contrastive self-supervised learning to handle this task. The overview of framework for our method is shown in Fig. 2, it contains two procedures: **Unpaired Domain Translation** and **Contrastive Self-supervised Learning**. For the former stage, we perform unpaired image-to-image translation between different domains and ForkGAN [19] is adopted to synthesize high-quality images due to its strong ability to perform disentanglement of domain-specific and domain-agnostic features. Here we only illustrate one direction (Domain $\mathbb{A} \rightarrow$ Domain \mathbb{B}) as an example, and the other direction (Domain $\mathbb{B} \rightarrow$ Domain \mathbb{A}) also performed the same operation. Given one image a from domain \mathbb{A} , we generate one reasonable image \tilde{b} in the target domain \mathbb{B} while preserving the content information of a . In the second stage, the random combination of five different transform operations (including the horizontal flipping, resizing, cropping, color jittering and grayscaling) are applied to a and \tilde{b} independently, yielding four different images: a^{t_1} , a^{t_2} , \tilde{b}^{t_3} and \tilde{b}^{t_4} , where a^{t_1} and a^{t_2} indicate the different transformed outputs of a after various transformations. To be noted, these two procedures are optimized separately.

Fig. 2 The overview of our proposed DaCo. Given an image a from domain \mathbb{A} , the Unpaired Domain Translation stage aims to generate reasonable image \tilde{b} in domain \mathbb{B} . The Contrastive Self-supervised Learning stage targets to minimize the within-domain distance and maximize the agreement between the cross-domain image pairs (a and \tilde{b}) through a self-generated soft constraint. Best viewed in color



3.2 Framework design

3.2.1 Unpaired domain translation

To achieve the domain translation between two visual domains: \mathbb{A} and \mathbb{B} (for example, nighttime and daytime domains), we perform the unpaired image-to-image translation between the two domains. Two reverse functions are introduced: translator $G_{\mathbb{A}}$ aims to generate $\tilde{b} = G_{\mathbb{A}}(a)$ in domain \mathbb{B} while translator $G_{\mathbb{B}}$ targets to synthesize the counterpart reconstruction of $\hat{a} = G_{\mathbb{B}}(\tilde{b})$ in domain \mathbb{A} . To make \tilde{b} as close to b as possible, the widely used adversarial loss is adopted:

$$\mathcal{L}_{adv}(G_{\mathbb{A}}, D_{\mathbb{B}}) = \mathbb{E}_b[\log D_{\mathbb{B}}(b)] + \mathbb{E}_{\tilde{b}}[\log(1 - D_{\mathbb{B}}(\tilde{b}))],$$

with $\tilde{b} = G_{\mathbb{A}}(a)$, (1)

where $D_{\mathbb{B}}$ is the domain-specific discriminator for domain \mathbb{B} . The reverse adversarial loss $\mathcal{L}_{adv}(G_{\mathbb{B}}, D_{\mathbb{A}})$ for the reverse direction is also computed, where $D_{\mathbb{A}}$ is the discriminator for domain \mathbb{A} . For the training stage, a and b are randomly selected, in other words, a and b are not paired. To preserve the content information after translation, the cycle-consistency loss \mathcal{L}_{cyc} [18] is adopted to link $G_{\mathbb{A}}$ and $G_{\mathbb{B}}$:

$$\mathcal{L}_{cyc}(G_{\mathbb{A}}, G_{\mathbb{B}}) = \mathbb{E}_{\mathbb{A}}[\|G_{\mathbb{B}}(G_{\mathbb{A}}(a)) - a\|_1] + \mathbb{E}_{\mathbb{B}}[\|G_{\mathbb{A}}(G_{\mathbb{B}}(b)) - b\|_1].$$

(2)

Through the pixel-wise distance, we can effectively preserve the content information after the unpaired image-to-image translation.

3.2.2 Contrastive self-supervised learning

To obtain the global content representations from images, we adopt one encoder $f(\cdot)$ to extract the learned feature representations. To make the learned feature efficient and meaningful, we firstly feed the four augmented images (a^{t_1} , a^{t_2} , \tilde{b}^{t_3} and \tilde{b}^{t_4}) into this encoder to obtain the features: $h_{a^{t_1}}$, $h_{a^{t_2}}$, $h_{\tilde{b}^{t_3}}$ and $h_{\tilde{b}^{t_4}}$, as shown in Fig. 2. Please note that \tilde{b} is generated from a through unpaired image-to-image translation, and the superscript t_1 , t_2 , t_3 and t_4 denote different transform combinations. Take a^{t_1} as an example, we adopt the widely used ResNet [14] to obtain $h_{a^{t_1}} = f(a^{t_1}) = \text{ResNet}(a^{t_1})$, where $h_{a^{t_1}} \in \mathbb{R}^d$ is the global feature after the average pooling layer. A small neural network projection head g is appended, it maps the global features to the output space where loss is computed. We apply a multiple layer perceptron architecture with one hidden layer to obtain $z_{a^{t_1}} = g(h_{a^{t_1}}) = W^{(2)}\sigma(W^{(1)}h_{a^{t_1}})$, where σ is a ReLU non-linearity function, $W^{(1)}$ and $W^{(2)}$ are the weights of fully connected layer. To be noted, the proposed framework allows various choices of the network architecture without any constraints.

For the training stage, we randomly sample a mini-batch of N examples and define the contrastive prediction task on

the augmented examples derived from the mini-batch, resulting in $4N$ data points. We do not sample negative examples explicitly. Instead, given a positive pair, we treat the other $4(N - 1)$ augmented examples within the mini-batch as negative examples. We compute two categories of losses: **within-domain** and **cross-domain** representation loss. Let $\text{sim}(u, v) = u^T v / \|u\| \|v\|$ denote the similarity between feature u and v , then the former loss for a positive pair within-domain of examples (a^{t1}, a^{t2}) is described as follows:

$$\mathcal{L}_{\text{within}}^{a^{t1}, a^{t2}} = -\log \frac{\exp(\frac{\text{sim}(z_{a^{t1}}, z_{a^{t2}})}{\tau})}{\sum_{x \in X} \exp(\frac{\text{sim}(z_{a^{t1}}, z_x)}{\tau})}, \quad (3)$$

where X is the set of augmented images in current mini-batch, and τ denotes a temperature parameter. This loss function is similar to some functions often used in metric learning, such as Arcface [40] and HASeparator [41]. The within-domain loss $\mathcal{L}_{\text{within}}^{a^{t1}, a^{t2}}$ is also computed. The latter cross-domain loss is a self-generated soft constraint, which is a weighted term controlled by hyper-parameter λ . And the cross-domain loss of examples (a^{t1}, b^{t3}) is described as:

$$\mathcal{L}_{\text{cross}}^{a^{t1}, b^{t3}} = -\log \frac{\exp(\frac{\text{sim}(z_{a^{t1}}, z_{b^{t3}})}{\tau})}{\sum_{x \neq a^{t1}} \exp(\frac{\text{sim}(z_{a^{t1}}, z_x)}{\tau})}, \quad (4)$$

similarly, the cross-domain loss $\mathcal{L}_{\text{cross}}^{a^{t1}, b^{t4}}$ is computed too. The final loss is computed across all positive pairs, it combines within-domain loss and cross-domain loss:

$$\mathcal{L} = \mathcal{L}_{\text{within}} + \lambda \mathcal{L}_{\text{cross}}, \quad (5)$$

where we set $\lambda = 0.8$ in all our experiments. More ablation studies about λ could be found in Section 4.5. Please note that there is no need to compute the cross-domain loss $\mathcal{L}_{\text{cross}}^{a^{t2}, b^{t3}}$ and $\mathcal{L}_{\text{cross}}^{a^{t2}, b^{t4}}$ explicitly, because those two constraints have potentially been considered by combining $\mathcal{L}_{\text{within}}^{a^{t1}, a^{t2}}$, $\mathcal{L}_{\text{cross}}^{a^{t1}, b^{t3}}$ and $\mathcal{L}_{\text{cross}}^{a^{t1}, b^{t4}}$. Such a design can drastically minimize the amount of computation and memory occupation while increasing training speed.

3.3 Why the combination?

The intuitive explanation of our DaCo is shown in Fig. 3, suppose one *anchor* image a in the nighttime domain, we perform different transform operations to a and obtain different transformed outputs: a^{t1} and a^{t2} , they all belong to the same instance. Suppose that the radius of the data transforms proposed in contrastive learning in feature space is θ (after $f(\cdot)$ and $g(\cdot)$ functions), and the distance between a and the corresponding daytime image b is β . Because the data augmentations we used are relatively simple, such as

random cropping, color jittering, random grayscaling and random flipping, the distribution of feature vectors obtained under different data augmentations in feature space is relatively centralized. However, images in different domains have great differences, and the distribution of their feature vectors is very scattered, we can observe $\beta \gg \theta$. After unpaired domain translation between nighttime and daytime domain, we can gain the *shifted anchor* \tilde{b} and the shifted distance is α . However, the distance between *anchor* and *shifted anchor* is still larger than θ , which is caused by the generation quality. To tightly align the translated domain and real domain, we conduct a self-generated soft constraint to reduce the distance. The self-generated supervision can indirectly help extract the domain-agnostic feature representations by checking whether the content information of the translated image is preserved. By computing this cross-domain consistency, we can better align the distribution of translated images and real images and weaken the negative impacts of the visual artifacts (The visual artifacts refer to unnatural textures, lines and grid phenomena, such as curved walls, smeared color patches in the sky, etc.).

4 Experiments and results

In this section, we first introduce the datasets, experimental setup and show some implementation details of our method. Then we compare our method with other methods on the three adopted datasets. Finally, some quantitative and qualitative experiments are shown to evaluate the effectiveness of all proposed components.

4.1 Datasets

The image translation/synthesis inevitably introduces noise and visual artifacts after the translation procedure. The unpaired translation setting poses further challenges. And our method focuses on obtaining image-level correspondences, so the widely used visual place recognition datasets (such as Aachen Day-Night [42], RobotCar [43], Mapillary [44] and CMU-Seasons [45] datasets) for obtaining the precise camera poses are not used in this paper. The adopted place recognition datasets are described below.

Tokyo 24/7 [26] dataset contains 4k high-resolution images taken at different times of the day by fixed webcams, provides different light conditions (day, sunset and night) for 125 places (100 for training and others for evaluation) under different viewing directions. In our experiments, we only choose day and night images at the same place as matched pairs, and images collected with different viewpoints are regarded as different instances. After the reorganization, we have 300 image pairs for training and 75 pairs for evaluation.

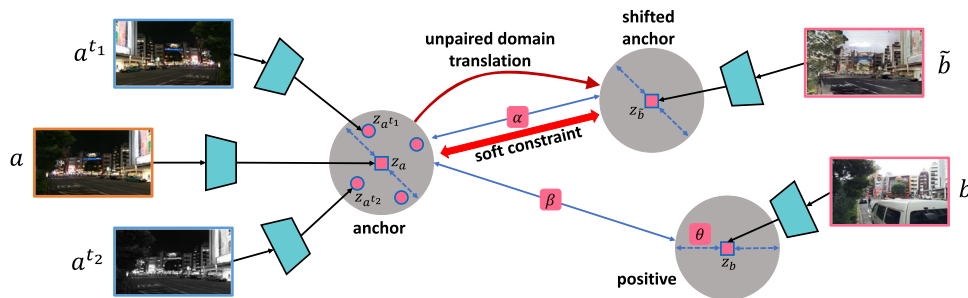


Fig. 3 The intuitive explanation of our DaCo. Suppose that the distance between the original *anchor* a in the nighttime domain and the *positive* b in the daytime domain is β without the unpaired domain translation. After the unpaired domain translation, we can obtain the *shifted anchor* \tilde{b} in the daytime domain, and the shifted distance α . We define that the radius of the transformations adopted in contrastive learning to be θ in the feature space. On one hand, we can observe $\beta \gg \theta$, so directly

applying the contrastive learning with various data augmentation techniques for cross-domain image retrieval leads to poor performance. On the other hand, after unpaired domain translation, though the distance between the *shifted anchor* and the *positive* is much smaller, there is still a large distance between *anchor* and *shifted anchor*. To alleviate this issue, our **DaCo** proposes a self-generated soft constraint to further align *anchor* and *shifted anchor*. Best viewed in color

Foggy cityscapes [46] is a recently proposed synthetic foggy dataset simulating fog on real scenes with three different levels of visibility: 150m, 300m and 600m visibility. For this dataset, we choose clear images and foggy images with 600m visibility following the original split. In all the experiments, 2975 clear and the corresponding foggy images are adopted for training and the other 500 clear and corresponding foggy images for evaluation. All the raw images are 2048×1024 in image resolution.

Synthia seqs is originally introduced in [47], which is captured in a synthetic engine. There are five different video sequences under various scenarios and traffic conditions with 5 FPS. We choose the sequences which contain sunset and rainy night images to perform cross-illumination and cross-weather image retrieval. After manually aligning the images from the two domains, we obtain 260 image pairs for training and 60 pairs for evaluation. These images are 1280×760 in image resolution.

4.2 Evaluation metric and experimental setup

Evaluation metric To evaluate the performance of different methods, we adopt Recall@K protocol [48] as the main evaluation metric. At the inference stage, the cosine similarities of feature vectors between the query image and other images in the test set are computed, and the top K images with the highest similarities are retrieved. A higher score indicates better retrieval performance.

Experimental setup There are a lot of unsupervised contrastive learning methods, here we only compare our DaCo with several representative and popular methods under the label-unknown setting (which means no matched image pairs are provided): NPID [32], SimCLR [12], MoCo [11] and SimSiam [37]. We also include the existing visual place

recognition algorithms (SIFT [27], BoW [28], VLAD [29] and NetVLAD [30]) for the sake of completeness. For all cross-domain visual place recognition evaluation, we perform evaluation in three settings:

1. The query image is from domain \mathbb{A} while the gallery images are from domain \mathbb{B} (termed $\mathbb{A} \rightarrow \mathbb{B}$).
2. The query image is from domain \mathbb{B} while the gallery images are from domain \mathbb{A} (termed $\mathbb{B} \rightarrow \mathbb{A}$).
3. The query image is from domain \mathbb{A} or domain \mathbb{B} while the gallery images are a mixture of images from domain \mathbb{A} and domain \mathbb{B} (termed $\mathbb{A} \leftrightarrow \mathbb{B}$).

4.3 Implementation details

For the unpaired domain translation at the first stage, we adopt ForkGAN [19] to perform unpaired image-to-image translation, and we set the image resolution as 1024×512 . After training, the synthesized images are resized to the original size through bicubic interpolation. To give an intuitive illustration, we have provided some cross-domain translation results in Fig. 4. Though the content representations of the translated images are well preserved, there are still some visual artifacts shown in the yellow boxes under the most challenging Sunset \rightarrow Rainy night setting, which could harm the image retrieval performance.

We adopt the same architecture of SimCLR [12] as backbone at the second stage. The projected output z is a 512-dimensional vector while the temperature τ inherited from [12] is 0.1. The number of the negative samples in NPID [32] and MoCo [11] is set to 4, 096 following the default setting. Besides, the momentum in MoCo [11] is 0.999 as the paper suggested, and the momentum in NPID [32] is 0.5. We set the image resolution as 256×256 , and the same data augmentations are adopted as SimCLR [12], such as random cropping,

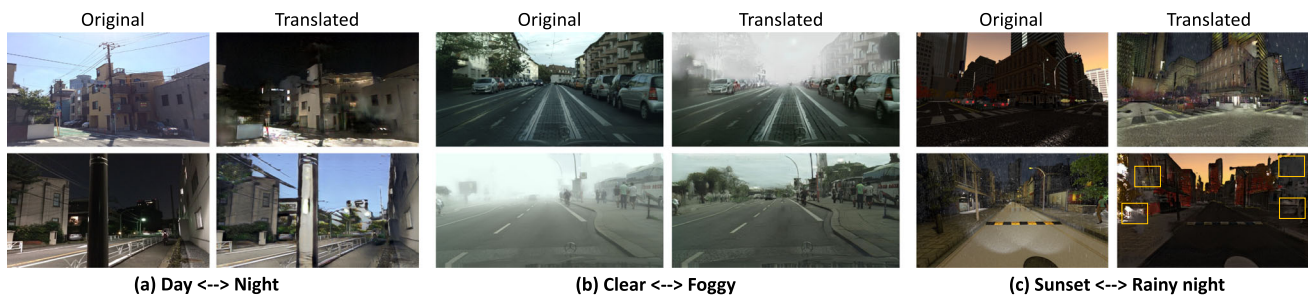


Fig. 4 The visual unpaired image-to-image translation results: (a) Day \leftrightarrow Night image translation; (b) Clear \leftrightarrow Foggy image translation; (c) Sunset \leftrightarrow Rainy night image translation. Best viewed in color and zoom-in. The regions covered by the yellow boxes show some visual artifacts

color jittering, random grayscaling and random flipping. All the methods have been optimized through the same iterations: 10,000. We choose Adam optimizer [49] with learning rate $1e-3$ and weight decay $1e-6$ to perform the optimization. Our framework is model-agnostic and we can choose various GAN architectures for image translation and contrastive learning methods for feature learning.

We then provide the detailed experimental setting to compare the proposed DaCo with previous SIFT [27], BoW [28], VLAD [29] and NetVLAD [30]. We first perform the feature matching between all the testing images from domain \mathbb{A} and \mathbb{B} and obtain all the feature matching pairs based on the brute-force manner. We obtain the top retrieved results based on the number of matched pairs. The image with the most matched pairs to the query image is regarded as the top-1 retrieved output. For BoW, we convert the extracted SIFT features to corresponding feature representations based on K-Means clustering then compute the similarity matrix for our cross-domain visual place recognition. For both VLAD and NetVLAD, the backbone is set to ResNet-50 for fair comparison and we optimize the whole model based on all the training images. The feature representations are computed for similarity computation. Our code is implemented with PyTorch [50] and our model is trained on a single TiTAN X GPU. The data split and the code are available on <https://github.com/leftthomas/DaCo>.

4.4 Comparison with other methods

Cross-illumination We target to perform the cross-illumination image retrieval task on the collected images with different illumination. The illumination disparity makes it difficult to identify whether the two images are from the same place. The experiments are performed on Tokyo 24/7 [26] dataset. The visual retrieval output (rank 1) is shown in Fig. 5 (a)-(c). From this figure, all the compared methods fail to obtain accurate visual place recognition under the three settings. The quantitative comparison is shown in Table 1, which also leads to the same conclusion. Even though SimCLR could achieve about 30% accuracy under Day \rightarrow Night and Night

\rightarrow Day settings, it obtained poor results under the more challenging Day \leftrightarrow Night setting. On the contrary, our DaCo can still obtain 45% R@1 accuracy, which is about 5 times as SimCLR. Both MoCo [11] and SimSiam [37] have the poor cross-domain image retrieval performance. We attribute these poor quantitative results to the different appearance between images from different domains, which lead to different gradient descent directions of the encoder. Thus, the two methods failed to obtain acceptable performance under this setting. And compared with the existing visual place recognition methods, our DaCo has achieved better performance. We observe that the learning-based NetVLAD [30] still suffer from the illumination changes and cannot obtain a very satisfactory performance. In contrast, the proposed method could achieve the most accurate cross-illumination matching.

Cross-weather image retrieval is also explored to show the effectiveness of the proposal. We conduct this task on the Foggy Cityscapes [46] dataset. The foggy makes some objects and buildings invisible. To perform robust visual place recognition, the model is supposed to possess a strong ability to extract efficient representations. The visual retrieval outputs are shown in Fig. 5 (d)-(f), the quantitative comparison of different methods is shown in Table 2. MoCo (0.4% and 0.4% for Clear \rightarrow Foggy and Foggy \rightarrow Clear, respectively) and SimSiam (0.6% and 0.8% for Clear \rightarrow Foggy and Foggy \rightarrow Clear, respectively) obtain unacceptable retrieval performance, which reflects that they cannot effectively extract domain-agnostic features. The poor performance can be attributed to mode collapse, which has been widely discussed in many researches about unsupervised contrastive learning, such as [34, 35, 37]. The two methods hardly eliminate the influence of domain-specific information (there is no constraint for the representation disentanglement), which leads to the failure to extract domain-agnostic features effectively. Our DaCo outperforms all the comparative unsupervised contrastive learning methods and most existing place recognition methods by a large margin. We observe that the SIFT [27] matching method could also achieve very competitive performance. However, it is extremely



Fig. 5 The visual image retrieval results of different methods with night settings: (a) Day \rightarrow Night; (b) Night \rightarrow Day; (c) Day \leftrightarrow Night; (d) Clear \rightarrow Foggy; (e) Foggy \rightarrow Clear; (f) Clear \leftrightarrow Foggy; (g) Sunset \rightarrow Rainy

night; (h) Rainy night \rightarrow Sunset; (i) Sunset \leftrightarrow Rainy night. The results with red boundary are failure cases, and the results with green boundary are success cases. Best viewed in color

Table 1 Cross-Illumination image retrieval performance on Tokyo 24/7 [26] dataset

Method	Day \rightarrow Night				Night \rightarrow Day				Day \leftrightarrow Night			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
SIFT [27]	30.67	40.00	53.33	60.00	36.00	38.67	54.67	65.33	14.67	24.67	36.00	46.00
BoW [28]	2.67	5.33	10.67	22.67	5.33	8.00	17.33	25.33	1.33	2.00	3.33	6.67
VLAD [29]	4.00	5.33	6.67	12.00	2.67	5.33	12.00	14.67	2.67	4.00	6.00	10.00
NetVLAD [30]	14.67	20.00	24.00	42.67	20.00	21.33	30.67	40.00	6.00	6.67	8.67	10.67
NPID [32]	6.67	10.67	17.33	28.00	8.00	12.00	16.00	21.33	2.00	4.00	5.33	10.67
MoCo [11]	5.33	6.67	12.00	17.33	6.67	9.33	12.00	16.00	0.00	0.00	0.00	0.67
SimCLR [12]	25.33	32.00	45.33	56.00	33.33	37.33	46.67	58.67	8.67	9.33	14.00	18.67
SimSiam [37]	4.00	5.33	9.33	16.00	4.00	5.33	6.67	14.67	1.33	1.33	1.33	3.33
DaCo	61.33	68.00	78.67	84.00	60.00	70.67	81.33	88.00	45.33	56.67	64.00	74.67

Score in bold indicates the best accuracy

Table 2 Cross-Weather image retrieval performance on Foggy Cityscapes [46] dataset

Method	Clear → Foggy				Foggy → Clear				Clear ↔ Foggy			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
SIFT [27]	99.60	99.60	99.80	99.80	99.60	99.60	99.60	99.80	97.70	99.10	99.50	99.60
BoW [28]	16.20	20.60	24.60	32.20	30.60	39.40	48.60	54.80	15.70	20.20	25.10	32.00
VLAD [29]	0.20	0.60	1.00	1.20	0.40	0.40	0.60	1.40	0.10	0.40	0.60	0.80
NetVLAD [30]	52.20	64.40	79.40	90.60	65.00	80.40	88.20	93.40	1.90	3.40	5.80	9.70
NPID [32]	5.60	7.80	10.20	17.20	5.20	8.40	12.60	19.80	0.20	0.50	0.70	1.00
MoCo [11]	0.40	0.80	1.40	3.00	0.40	0.80	1.80	3.40	0.20	0.20	0.20	0.20
SimCLR [12]	34.60	48.00	59.60	71.40	54.60	66.80	78.40	86.00	0.50	1.00	1.70	2.90
SimSiam [37]	0.60	1.00	1.60	2.40	0.80	1.00	1.80	2.80	0.00	0.00	0.00	0.00
DaCo	98.80	99.60	99.80	100.0	98.60	99.20	99.40	99.80	90.20	96.00	98.60	99.20

Score in bold indicates the best accuracy

time-consuming to perform the brute-force feature matching and the time cost increases exponentially with the number of testing images. Furthermore, our method can obtain 90.2% R@1 precision under the challenging Clear ↔ Foggy setting. It also shows that the features extracted by our method are robust to the environmental condition changes.

Cross-illumination & weather We design a joint cross illumination and weather image retrieval on a much more challenging Synthia Seqs [47] dataset. Besides the low illumination, the raindrops and the reflection of various lights result in extreme difficulty implementing precise visual place recognition. The visual quantitative comparison of different methods is shown in Fig. 5 (g)-(i) while the quantitative results in Table 3. Our DaCo has achieved the best retrieval precision among all the evaluation metrics, while other comparative methods fail to obtain acceptable results (R@1 accuracies are all below 26%). The accuracies of comparative methods are very poor for the most challenging setting: Sunset ↔ Rainy night, which imply the feature disentanglement is not well performed by these methods. This also

shows that the constraints we proposed are very helpful to the results.

4.5 Ablation studies

Distribution visualization To explore the reason why compared methods fail to perform the challenging cross-domain image retrieval, we visualized the test sample distribution by T-SNE based on the learned global feature representation. The sample distribution of the above three tasks is illustrated in Fig. 6. From this figure, NPID, MoCo, SimCLR and SimSiam failed to obtain the domain-agnostic feature representations: the distribution of the samples from two different domains is completely separated, as for NPID, although the distribution for Tokyo 24/7 and Synthia Seqs datasets is mixed up, the total number of correct retrieved is very low, which means that it is unable to successfully learn the correspondence between two domains. Thus, these methods cannot directly conduct cross-domain image retrieval when the domain gap is huge. The dramatic performance drop

Table 3 Cross-Illumination&Weather image retrieval performance on Synthia Seqs [47] dataset

Method	Sunset → Rainy night				Rainy night → Sunset				Sunset ↔ Rainy night			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
SIFT [27]	30.00	46.67	56.67	66.67	5.00	13.33	25.00	45.00	2.50	6.67	15.83	30.83
BoW [28]	5.00	11.67	26.67	36.67	8.33	13.33	21.67	35.00	2.50	5.00	8.33	13.33
VLAD [29]	5.00	6.67	6.67	13.33	5.00	5.00	8.33	15.00	3.33	5.00	7.50	8.33
NetVLAD [30]	18.33	30.00	45.00	58.33	28.33	35.00	40.00	53.33	5.00	9.17	15.00	26.67
NPID [32]	8.33	10.00	11.67	20.00	6.67	6.67	18.33	21.67	5.00	5.00	10.00	11.67
MoCo [11]	5.00	5.00	8.33	16.67	10.00	15.00	16.67	25.00	1.67	1.67	1.67	3.33
SimCLR [12]	23.33	30.00	46.67	51.67	25.00	33.33	48.33	63.33	6.67	10.83	16.67	20.83
SimSiam [37]	10.00	13.33	15.00	23.33	5.00	10.00	21.67	30.00	3.33	5.00	6.67	8.33
DaCo	55.00	68.33	75.00	88.33	41.67	51.67	70.00	83.33	25.00	34.17	50.00	66.67

Score in bold indicates the best accuracy

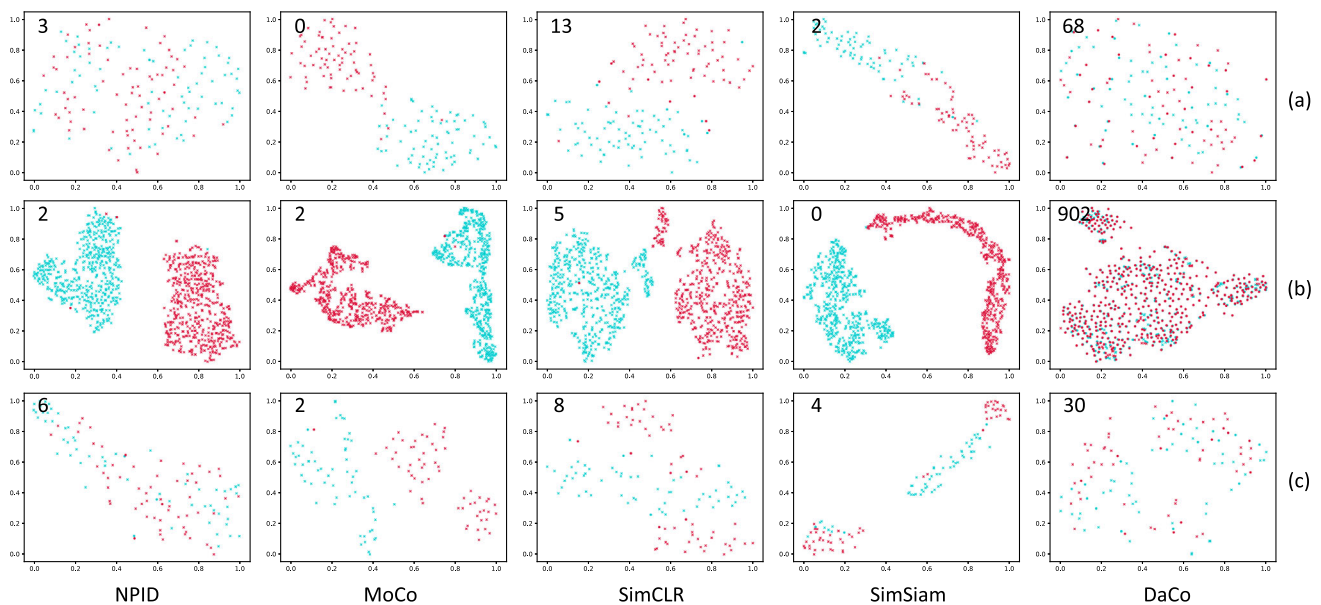


Fig. 6 The feature distribution visualization of various methods by using T-SNE on these datasets: (a) Tokyo 24/7 [26] dataset; (b) Foggy Cityscapes [46] dataset; (c) Synthia Seqs [47] dataset. The cyan-blue and red color indicate samples from two different domains, respectively. The samples with circle shape means the retrieved image with highest

similarities is correct, on the contrary, the samples with cross shape means the retrieved image with highest similarities is wrong. The total number of correct retrieved is shown in the top left of each sub-figure. Best viewed in color and zoom-in

under the third setting $\mathbb{A} \leftrightarrow \mathbb{B}$ can also be blamed for this separation. On the contrary, our method can mix up the distribution of the samples from two domains, which indicates the domain-agnostic representations can be perfectly extracted.

Hyper-parameter selection We have also explored the influence of hyper-parameter λ on Foggy Cityscapes [46] dataset. Extensive values of λ are chosen, and all the quantitative comparisons of different settings are shown in Table 4. As observed, we could see the cross-domain image retrieval per-

formance becomes better with the increase of λ . For the self-generated soft constraint, we can control the distance between the real domain and the translated domain by changing the value of λ . When $\lambda = 0.8$, our DaCo achieved the best retrieval performance under most of the evaluation metrics (R@1 accuracy is given more priority in visual place recognition).

If we follow the pipeline architecture (“GAN + Recognition”) proposed in [19, 20], which means without the soft

Table 4 Cross-Weather image retrieval performance of different hyper-parameter λ on Foggy Cityscapes [46] dataset

λ	Clear \rightarrow Foggy				Foggy \rightarrow Clear				Clear \leftrightarrow Foggy			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
0.0	42.80	54.00	67.40	79.80	42.20	57.80	69.20	80.60	2.70	4.00	6.90	11.50
0.1	42.80	56.00	70.20	81.60	65.20	84.00	95.20	98.20	2.10	3.80	6.10	8.40
0.2	64.20	76.00	85.60	91.40	78.00	89.20	96.00	98.80	5.60	8.90	12.80	18.40
0.3	60.40	70.80	80.80	88.80	85.60	91.80	97.20	98.80	5.30	8.30	12.00	17.40
0.4	93.40	96.60	98.60	99.00	93.60	96.20	98.00	99.40	77.00	85.70	91.50	94.50
0.5	95.60	98.00	99.00	99.80	96.60	99.00	99.60	99.80	82.40	89.60	94.40	97.10
0.6	90.20	95.40	98.20	99.60	93.60	98.00	99.20	99.80	31.10	44.00	59.00	74.70
0.7	95.60	98.20	99.20	99.80	96.60	99.00	99.40	100.0	82.40	91.30	95.90	98.10
0.8	98.80	99.60	99.80	100.0	98.60	99.20	99.40	99.80	90.20	96.00	98.60	99.20
0.9	94.60	97.80	98.80	99.40	96.60	99.40	99.60	100.0	78.60	88.80	94.60	97.80
1.0	93.60	96.40	98.80	99.60	94.00	97.00	98.40	99.40	78.10	87.10	93.50	95.90

Score in bold indicates the best accuracy

constraint (when $\lambda = 0$), the unpaired image-to-image translation could be regarded as an effective data transformation and leads some performance gain compared with the alternative unsupervised methods. However, GAN could introduce the noise and uncertainty and there is still minor shift between real and generated samples. To further promote the performance, we design a self-generated soft constraint to better align the real sample distribution and generated sample distribution. Our DaCo utilizes an appropriate value of λ and has achieved a huge performance improvement. Take a step furthermore, if λ is too small (e.g., $\lambda = 0.2$), we cannot obtain effective domain-agnostic content representations through the alignment. In contrast, $\lambda = 1.0$ leads to worse performance compared with the setting of $\lambda = 0.8$. We attribute this phenomenon to the influence of balance between the within-domain loss and cross-domain loss. Our appropriate weighted self-generated soft constraint could effectively alleviate the negative influences, this is also why we call that “soft” constraint.

The impact of additional cross-domain loss Recalling what we mentioned in Section 3.2.2, we only consider the cross-domain loss $\mathcal{L}_{cross}^{a^1, \tilde{b}^3}$ and $\mathcal{L}_{cross}^{a^1, \tilde{b}^4}$, ignore $\mathcal{L}_{cross}^{a^2, \tilde{b}^3}$ and $\mathcal{L}_{cross}^{a^2, \tilde{b}^4}$. Because we think those two constraints ($\mathcal{L}_{cross}^{a^2, \tilde{b}^3}$ and $\mathcal{L}_{cross}^{a^2, \tilde{b}^4}$) have potentially been considered by combining $\mathcal{L}_{within}^{a^1, a^2}$, $\mathcal{L}_{cross}^{a^1, \tilde{b}^3}$ and $\mathcal{L}_{cross}^{a^1, \tilde{b}^4}$. To verify this conclusion, we have conducted experiments on Foggy Cityscapes [46] dataset and the quantitative results are reported in Table 5. As demonstrated, we can observe that the performance improvement based on these two constraints is limited (e.g., in the case of Clear

→ Foggy, R@1 only increase 0.2%). However, introducing these two constraints inevitably results in more computation costs and noticeable memory burden. To reduce the computational cost and alleviate memory consumption, we did not include these two loss constraints in our method considering the tradeoff between performance and computational cost.

The impact of data augmentations The data augmentation strategy is very important for self-supervised learning. Our DaCo adopts the same data augmentation strategy as SimCLR [12] (including random cropping, color jittering, random grayscaling and random flipping) for a fair comparison. We have also explored the influence of choosing different data augmentation strategies on Foggy Cityscapes [46] dataset. The experimental results are shown in Table 6. The random cropping is used as our baseline (refer to the first row in Table 6). Applying the color jittering or grayscaling alone can greatly improve the performance, while solely using the horizontal flipping leads to performance degradation. However, when any two data augmentation strategies are combined, the performance can be further improved, especially for the case of Clear ↔ Foggy (e.g., the performance of R@1 for color jittering+horizontal flipping vs. color jittering vs. horizontal flipping is 79.80 vs. 38.60 vs. 23.60). We can obtain the best performance by combining all three data augmentation strategies together. It can be seen that an appropriate choice of data augmentation is crucial to the final performance.

Failed cases analysis To comprehensively evaluate the proposed method, we also show the failed cases in Fig. 7 to

Table 5 Ablation studies for the impact of additional cross-domain loss ($\mathcal{L}_{cross}^{a^2, \tilde{b}^3}$ and $\mathcal{L}_{cross}^{a^2, \tilde{b}^4}$) on Foggy Cityscapes [46] dataset

$\mathcal{L}_{cross}^{a^2, \tilde{b}^3} + \mathcal{L}_{cross}^{a^2, \tilde{b}^4}$	Clear → Foggy				Foggy → Clear				Clear ↔ Foggy			
	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
–	98.80	99.60	99.80	100.0	98.60	99.20	99.40	99.80	90.20	96.00	98.60	99.20
✓	99.00	99.80	100.0	100.0	98.20	99.40	100.0	100.0	92.10	97.00	98.70	99.80

Score in bold indicates the best accuracy

Table 6 Ablation studies for the impact of different data augmentations on Foggy Cityscapes [46] dataset

Color Jittering	Gray Scaling	Horizontal Flipping	Clear → Foggy				Foggy → Clear				Clear ↔ Foggy			
			R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8	R@1	R@2	R@4	R@8
–	–	–	62.60	75.00	82.00	88.40	73.60	82.60	87.80	93.00	25.70	37.40	51.80	64.30
✓	–	–	89.00	94.00	99.00	99.40	94.80	98.00	99.60	100.0	38.60	55.70	72.00	84.00
–	✓	–	83.00	91.20	95.60	97.60	82.40	90.40	93.60	96.80	56.00	67.30	79.50	89.10
–	–	✓	61.40	74.20	84.00	91.40	65.00	77.80	86.00	92.20	23.60	33.20	46.00	57.60
✓	✓	–	90.80	96.20	98.80	99.40	92.20	98.00	98.80	99.40	76.20	85.20	93.30	96.90
✓	–	✓	92.00	96.80	98.80	99.60	94.60	99.20	99.80	100.0	79.80	89.10	95.30	98.10
–	✓	✓	75.20	85.80	93.20	97.40	69.80	81.80	88.20	94.00	36.10	49.30	60.50	72.50
✓	✓	✓	98.80	99.60	99.80	100.0	98.60	99.20	99.40	99.80	90.20	96.00	98.60	99.20

Score in bold indicates the best accuracy



Fig. 7 The failed image retrieval cases of DaCo with all settings on three datasets: (a) Tokyo 24/7 [26]; (b) Foggy Cityscapes [46]; (c) Synthia Seqs [47]. By comparing the query, ground truth and retrieved images, we can find that there are many similar parts between these

images. Many of these images are only slightly different, which indirectly explains why our method misclassifies these cases. Best viewed in color and zoom-in

enable us to know the proposed method's limitations. Take the images of Day \rightarrow Night as an example, no matter the query image or retrieved image, they all contain traffic lights and pedestrian crosswalk. Further more, their scene structure is also very similar, such as a row of trees grew behind the traffic lights. The only difference between them is the location of building. This also helps us understand why our method has achieved such results. Take the images of Rainy night \rightarrow Sunset as another example, the buildings of query image and retrieved image have the same obliquity, and the architectural style is very similar, as for the ground truth image, it has the different obliquity to the query image, but has the same architectural style. This example enlightens us whether we can consider incorporating deformation into our method to further improve the robustness of image retrieval. For other cases in Fig. 7, the query image and retrieved image have similar parts to some extent, this also shows that our method effectively extracts some cross-domain features.

5 Conclusion

In this paper, we propose a novel domain-agnostic contrastive learning method termed “DaCo” to perform visual place recognition based on cross-domain image retrieval without matched image pairs. To tightly integrate the domain-agnostic feature representations and self-supervision, we design a soft constraint to boost the performance of the challenging cross-domain (especially the cross-illumination and cross-weather) image retrieval task. The impressive results and the annotation-free training scheme have demonstrated a high potential to be applied in a real-world environment, such as autonomous driving and robotics. Since our method focuses on obtaining image-level correspondence rather than accurate geographic positioning, it is still unable to solve the problem of geographic positioning at this stage. In the future, we will consider integrating geographic verification [24] and

tensor decomposition [51, 52] to further expand the application scope of this method.

Author Contributions All authors contributed to the study's conception and design. Material preparation, data collection and analysis were performed by Hao Ren. The first draft of the manuscript was written by Ziqiang Zheng and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by Scientific and Technological innovation action plan of Shanghai Science and Technology Committee (No.22511102202), Fudan University Double First-class Construction Fund (No. XM03211178).

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Zaffar M, Garg S, Milford M, Kooij J, Flynn D, McDonald-Maier K, Ehsan S (2021) Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable view-point and appearance change. *International Journal of Computer Vision* 129(7):2136–2174
2. Özdemir A, Scerri M, Barron AB, Philippides A, Mangan M, Vasiliaki E, Manneschi L (2022) Echovpr: Echo state networks for visual place recognition. *IEEE Robotics and Automation Letters* 7(2):4520–4527
3. Thoma, J., Paudel, D.P., Gool, L.V.: Soft contrastive learning for visual localization. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 11119–11130 (2020)
4. Skrzypczyński, P.: Mobile robot localization: Where we are and what are the challenges? *International Conference Automation*, 249–267 (2017)
5. Li, L., Kong, X., Zhao, X., Huang, T., Li, W., Wen, F., Zhang, H., Liu, Y.: Ssc: Semantic scan context for large-scale place recognition. In: *IEEE RSJ International Conference on Intelligent Robots and Systems*, pp. 2092–2099 (2021)
6. Wang, H., Pi, J., Qin, T., Shen, S., Shi, B.E.: Slam-based localization of 3d gaze using a mobile eye tracker. In: *ACM Symposium on Eye Tracking Research & Applications*, p. 65 (2018)

7. Fine-tuning cnn image retrieval with no human annotation
8. Zheng L, Yang Y, Tian Q (2018) Sift meets cnn: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(5):1224–1244
9. Gadd, M., De Martini, D., Newman, P.: Contrastive learning for unsupervised radar place recognition. In: *International Conference on Advanced Robotics*, pp. 344–349 (2021)
10. Jaiswal A, Babu AR, Zadeh MZ, Banerjee D, Makedon F (2020) A survey on contrastive self-supervised learning. *Technologies* 9(1):2
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738 (2020)
12. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning*, vol. 1, pp. 1597–1607 (2020)
13. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 22243–22255 (2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
15. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2011–2023 (2018)
16. Zhao, S., Yue, X., Zhang, S., Li, B., Zhao, H., Wu, B., Krishna, R., Gonzalez, J.E., Sangiovanni-Vincentelli, A.L., Seshia, S.A., Keutzer, K.: A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks*, 1–21 (2020)
17. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville AC, Bengio Y (2020) Generative adversarial networks. *Communications of The ACM* 63(11):187–208
18. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *IEEE International Conference on Computer Vision*, pp. 2242–2251 (2017)
19. Zheng, Z., Wu, Y., Han, X., Shi, J.: Forkgan: Seeing into the rainy night. In: *European Conference on Computer Vision*, pp. 155–170 (2020)
20. Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., Gool, L.V.: Night-to-day image translation for retrieval-based localization. In: *International Conference on Robotics and Automation*, pp. 5958–5964 (2019)
21. Lee, K., Zhu, Y., Sohn, K., Li, C.-L., Shin, J., Lee, H.: i-mix: A domain-agnostic strategy for contrastive representation learning. In: *International Conference on Learning Representations* (2021)
22. Verma, V., Luong, M.-T., Kawaguchi, K., Pham, H., Le, Q.V.: Towards domain-agnostic contrastive learning. In: *International Conference on Machine Learning*, vol. 139, pp. 10530–10541 (2021)
23. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: Beyond empirical risk minimization. In: *International Conference on Learning Representations* (2017)
24. Chang, C., Yu, G., Liu, C., Volkovs, M.: Explore-exploit graph traversal for image retrieval. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9423–9431 (2019)
25. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14141–14152 (2021)
26. Akihiko, T., Relja, A., Josef, S., Masatoshi, O., Tomas, P.: 24/7 place recognition by view synthesis. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817 (2015)
27. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110
28. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2007)
29. Arandjelovic, R., Zisserman, A.: All about vlad. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1578–1585 (2013)
30. Arandjelovic R, Gronat P, Torii A, Pajdla T, Sivic J (2018) Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(6):1437–1451
31. Liu, L., Li, H., Dai, Y.: Stochastic attraction-repulsion embedding for large scale image localization. In: *IEEE International Conference on Computer Vision*, pp. 2570–2579 (2019)
32. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742 (2018)
33. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: *European Conference on Computer Vision*, pp. 776–794 (2019)
34. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924 (2020)
35. Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284 (2020)
36. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In: *IEEE International Conference on Computer Vision*, pp. 9588–9597 (2021)
37. Chen, X., He, K.: Exploring simple siamese representation learning. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2021)
38. Liu, M.-Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. In: *IEEE International Conference on Computer Vision*, pp. 10551–10560 (2019)
39. Bhattacharjee, D., Kim, S., Vizier, G., Salzmann, M.: Dunit: Detection-based unsupervised image-to-image translation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4787–4796 (2020)
40. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699 (2019)
41. Kansizoglou, I., Santavas, N., Bampis, L., Gasteratos, A.: Haseparator: Hyperplane-assisted softmax. In: *IEEE International Conference on Machine Learning and Applications*, pp. 519–526 (2020)
42. Sattler, T., Weyand, T., Leibe, B., Kobbelt, L.: Image retrieval for image-based localization revisited. In: *British Machine Vision Conference*, vol. 1, p. 4 (2012)
43. Maddern W, Pascoe G, Linegar C, Newman P (2017) 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research* 36(1):3–15
44. Jafarzadeh, A., Antequera, M.L., Gargallo, P., Kuang, Y., Toft, C., Kahl, F., Sattler, T.: Crowddriven: A new challenging dataset for outdoor visual localization. In: *IEEE International Conference on Computer Vision*, pp. 9845–9855 (2021)
45. Bansal, A., Badino, H., Huber, D.: Understanding how camera configuration and environmental conditions affect appearance-based

- localization. In: IEEE Intelligent Vehicles Symposium Proceedings, pp. 800–807 (2014)
46. Sakaridis, C., Dai, D., Hecker, S., Gool, L.V.: Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In: European Conference on Computer Vision, pp. 707–724 (2018)
 47. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3234–3243 (2016)
 48. Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4004–4012 (2016)
 49. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
 50. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, vol. 32, pp. 8026–8037 (2019)
 51. Hu C, Wang Y, Gu J (2020) Cross-domain intelligent fault classification of bearings based on tensor-aligned invariant subspace learning and two-dimensional convolutional neural networks. *Knowledge-Based Systems* 209:106214
 52. Hu C, He S, Wang Y (2021) A classification method to detect faults in a rotating machinery based on kernelled support tensor machine and multilinear principal component analysis. *Applied Intelligence* 51(4):2609–2621

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



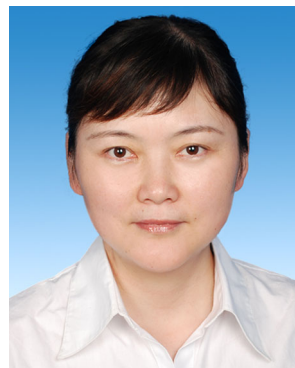
Hao Ren received his B.Eng. degree in software engineering from the Zhejiang University of Technology in 2017. He is currently with the Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai, China. His research interests include pattern recognition and computer vision.



Ziqiang Zheng received his B.Eng. degree in communication engineering from the Ocean University of China in 2019. After his graduation, he joined UISEE. His research interests include generative adversarial network and computer vision.



Yang Wu received a BS degree and a Ph.D. degree from Xi'an Jiaotong University in 2004 and 2010, respectively. He is currently a principal researcher with Tencent AI Lab. From Jul. 2019 to May 2021, he was a program-specific senior lecturer with the Department of Intelligence Science and Technology, Kyoto University. He was an assistant professor of the NAIST International Collaborative Laboratory for Robotics Vision, Nara Institute of Science and Technology (NAIST), from Dec. 2014 to Jun. 2019. From 2011 to 2014, he was a program-specific researcher with the Academic Center for Computing and Media Studies, Kyoto University. His research is in the fields of computer vision, pattern recognition, as well as multimedia content analysis, enhancement and generation.



Hong Lu received the B.Eng. and M.Eng. degrees in computer science and technology from Xidian University, Xi'an, China, in 1993 and 1998, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2005. From 1993 to 2000, she was a Lecturer and a Researcher with the School of Computer Science and Technology, Xidian University. From 2000 to 2003, she was a Research Student with the School of Electrical and Electronic Engineering, Nanyang Technological University. Since 2004, she has been with the School of Computer Science, Fudan University, Shanghai, China, where she is currently a Professor. Her current research interests include computer vision, machine learning, pattern recognition, and robotic tasks.