

# VG-SSL: Benchmarking Self-supervised Representation Learning Approaches for Visual Geo-localization

Jiuhong Xiao  
New York University  
jx1190@nyu.edu

Gao Zhu  
New York University  
gz706@nyu.com

Giuseppe Loianno  
New York University  
loianno@nyu.edu

## Abstract

*Visual Geo-localization (VG) is a critical research area for identifying geo-locations from visual inputs, particularly in autonomous navigation for robotics and vehicles. Current VG methods often learn feature extractors from geo-labeled images to create dense, geographically relevant representations. Recent advances in Self-Supervised Learning (SSL) have demonstrated its capability to achieve performance on par with supervised techniques with unlabeled images. This study presents a novel VG-SSL framework, designed for versatile integration and benchmarking of diverse SSL methods for representation learning in VG, featuring a unique geo-related pair strategy, **GeoPair**. Through extensive performance analysis, we adapt SSL techniques to improve VG on datasets from hand-held and car-mounted cameras used in robotics and autonomous vehicles. Our results show that contrastive learning and information maximization methods yield superior geo-specific representation quality, matching or surpassing the performance of state-of-the-art VG techniques. To our knowledge, This is the first benchmarking study of SSL in VG, highlighting its potential in enhancing geo-specific visual representations for robotics and autonomous vehicles. The code is publicly available at <https://github.com/arplaboratory/VG-SSL>.*

## 1. Introduction

Visual Geo-localization (VG), also known as Visual Place Recognition, is a rapidly growing research area in computer vision [2, 4, 8, 9] and robotics [11, 16, 21, 51]. Its primary goal is to identify geographical locations based on visual features. Most VG methods follow the image retrieval paradigm, building a database of robust, compact representations (e.g., embeddings or histograms of orientation [32]) from geo-referenced images containing longitude and latitude data. Recent approaches heavily rely on deep neural networks [30] to create these representations,

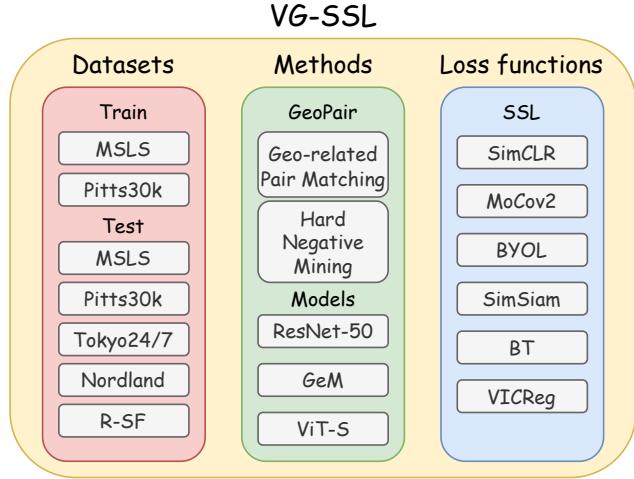


Figure 1. **VG-SSL** Framework Overview: This framework integrates various Visual Geo-localization (VG) datasets, models, and Self-Supervised Learning (SSL) loss functions for representation learning. It benchmarks VG performance across different SSL strategies trained with the geo-related pair strategy, **GeoPair**, and offers an in-depth analysis of SSL method settings tailored for geo-specific representation learning.

projecting images into an embedding space. The representation of a query image is then matched to the closest embedding in the database to retrieve geo-referenced information. For accurate matching, the feature extractor must produce representations that are resilient to noise, viewpoint changes, and dynamic objects. This aligns with the principles of representation learning, which focuses on generating robust and informative data representations, with or without specific downstream tasks. This concept has become increasingly important in computer vision, natural language processing [7], and robotics [39].

Self-Supervised Learning (SSL) techniques recently emerged as popular solutions for representation learning in computer vision. These methods rely on pre-training feature extractors on pretext tasks [34] with vast amounts of

unlabeled data to learn visual representation. Once pre-training, these models can be fine-tuned by freezing the pre-trained feature extractor and training additional downstream heads for specific tasks. With a pre-trained feature extractor, they can achieve on-par performance with much less labeled data. For example, a common SSL training strategy is contrastive learning [13, 14, 26]. This approach aims to minimize the embedding distance between augmented image views of the same instance and maximize the distance between views of different instances.

In pursuit of the shared objective of VG and SSL methods to learn robust and recognizable representations, this study aims to leverage SSL methodologies for VG tasks to assess their capability in learning geo-specific representations. However, relying solely on SSL’s data augmentation is insufficient for a comprehensive understanding of real-world geographic relationships, which can be enhanced by geo-labels. To bridge this gap, we introduce a novel pairing strategy that incorporates both data augmentation and geo-labels within SSL methods, aiming to improve the learning of geo-specific representations significantly.

Our main contributions can be summarized as follows:

- We introduce the novel **VG-SSL** framework (Fig. 1) to benchmark self-supervised representation learning methods such as SimCLR, MoCov2, BYOL, SimSiam, Barlow Twins, and VICReg. Validated across multiple large-scale VG datasets, this framework offers a comprehensive view of SSL methods in VG contexts and also includes a unified interface for extending to other SSL approaches.
- We propose the **GeoPair** strategy as a new data pairing strategy for VG tasks. It not only leverages geo-labels to learn and capture geographical relationships as well as employs data augmentation and advanced SSL loss functions to enhance sample efficiency.
- Extensive evaluations of the proposed framework on robotics and autonomous vehicle VG datasets demonstrate that contrastive learning and information maximization methods effectively learn geo-specific representations with complex spatial relationships of data, achieving performance on par with or superior to current leading VG methods. Further analysis identifies key architectural settings that significantly impact VG performance.

To our knowledge, this is the first unified framework for benchmarking SSL methods in VG tasks. It not only lays the foundation for enhancing geo-localization with SSL techniques but also provides an effective training approach for SSL visual foundational models to learn geo-relationships in data.

## 2. Related Works

**Representation Learning for VG.** VG methods aim to extract robust geo-related representations from images to differentiate distinct locations. Traditional VG approaches [17, 18, 24, 33] rely on handcrafted templates, which are generally less resilient to noise and appearance changes (e.g., illumination). In contrast, recent deep learning-based VG methods [2–4, 8, 53, 55] leverage deep neural networks [30] for geo-specific representation learning, incorporating local feature extraction and trainable feature aggregation modules. Popular feature aggregation techniques include MAC [37], GeM pooling [36], and NetVLAD [4], with various extensions [25, 28, 41] widely adopted. Most VG methods employ *metric learning* with triplet, pair, or multi-similarity [48] loss for distinguishing positive samples from the same place and negative samples from different places. A recent benchmark [9] highlights the effectiveness of triplet loss with Hard Negative Mining (HNM). Additionally, a trend toward classification-based VG methods [8, 10, 46] is emerging, grouping images by location and distinguishing between different classes.

VG methods can also be classified into one-stage approaches [4, 8, 9, 22, 31, 36], which use global image embeddings, and two-stage approaches [25, 38, 47, 56], which enhance accuracy by generating local features and re-ranking top candidates. While two-stage methods offer better performance, they require more memory and longer inference times. Compared with these works, our work focuses on benchmarking SSL approaches using geo-referenced imagery for geo-specific representation learning, establishing an SSL baseline compared to state-of-the-art VG methods.

**Self-supervised Representation Learning.** Recent advances in self-supervised representation learning have achieved performance on par with, or even surpassing, supervised methods in computer vision tasks. These methods learn robust visual representations from unlabeled image data through advanced data augmentation. The key challenge in SSL is avoiding model collapse, where the model produces identical outputs regardless of input, making it difficult to differentiate samples. We focus on three categories of SSL methods: *Contrastive Learning* [13, 14, 26], *Self-distillation Learning* [15, 23], and *Information Maximization* [6, 54]. A detailed architecture comparison is provided in Sec.D of the appendix.

**Contrastive Learning.** Methods like SimCLR [13] and MoCov2 [14] align embeddings of augmented views of the same image while maximizing the embedding distance from unrelated images. To prevent collapse, SimCLR uses a large batch size and a projection head, while MoCo [26] and MoCov2 introduce InfoNCE [35] loss and a momentum encoder, maintaining a queue of embeddings to avoid collapse without requiring large batch sizes.

**Self-distillation Learning.** Methods like BYOL [23]

and SimSiam [15] prevent collapse using a teacher-student framework. BYOL uses a momentum target encoder and cosine distance loss between predicted and actual embeddings. SimSiam highlights the importance of stop-gradient operations and batch normalization in avoiding collapse.

*Information maximization.* Methods like Barlow Twins [54] and VICReg [6] decorrelate embeddings to maximize information. Barlow Twins uses a cross-correlation matrix to suppress off-diagonal elements, while VICReg employs Variance-Invariance-Covariance (VIC) Regularization to maintain variance and reduce covariance between embedding elements.

**Intersection between VG and SSL.** Recent works have explored the intersection of VG and SSL methods. For instance, [49] reviews SSL techniques for remote sensing classification within a pretrain-finetune framework. [5] employs the MoCov2 [14] framework with a memory dictionary, using geographical information prediction as a pretext task. [1] improves hard negative mining in VG through a MoCo-like memory bank. Unlike these studies, our work focuses on a benchmarking framework for SSL methods in the VG context, introducing the GeoPair strategy. GeoPair links VG metric learning with SSL by using geo-labels to capture real geo-relationships and employing SSL techniques to enhance sample efficiency by data augmentations.

### 3. Methodology

In our VG-SSL framework, depicted in Fig. 2, we have designed the GeoPair strategy to build image pairs to accommodate multiple SSL training strategies. This strategy provides query images ( $I_q$ ), positive images ( $I_{k^p}$ ), and augmented negative images ( $I_{k^n}^t$  and  $I_{k^n}^{t'}$ ) for the feature extractor  $F$ . The extractor comprises a local feature extractor  $F_l$  (e.g., ResNet-50 [27]), a global feature aggregation module  $F_g$  (e.g., GeM [36]), and a projection head  $F_p$  (an MLP with consistent hidden and output dimensionality), inspired by [13], to map feature embeddings into a projected embedding space. The projection head is a common component in SSL methods for enhancing embedding quality. Subsequently, we design a unified module that allows the selection of various SSL loss functions to utilize these projected embeddings for training  $F$ .

We denote  $D_g$  as the dimension of feature embeddings output by  $F_g$ ,  $D_p$  as the dimension of projected embeddings output by  $F_p$ ,  $N$  as the batch size,  $L$  as the number of layers in the projection head,  $q$  as the projected embeddings of  $I_q$ ,  $k^p$  and  $k^n$  as the projected embeddings of  $I_{k^p}$  and  $I_{k^n}$ , respectively. The terms  $q_b = [q_{b,1} \ q_{b,2} \ \cdots \ q_{b,D_p}]^\top$  and  $k_b^p = [k_{b,1}^p \ k_{b,2}^p \ \cdots \ k_{b,D_p}^p]^\top$  are the  $b$ -th embedding with dimension  $D_p$  of queries and positives respectively. The terms  $q = [q_1 \ q_2 \ \cdots \ q_N]$  and  $k^p = [k_1^p \ k_2^p \ \cdots \ k_N^p]$  are the batch of paired embed-

dings for training. Optionally, if we choose to use negative samples in training, we employ Hard Negative Mining (HNM) to select negatives. Our preliminary results show that the model trained with randomly selected negatives yields performance similar to that of those trained with only query-positive pairs. We use data augmentation  $t, t' \sim T$  to augment  $I_{k^n}$  and get augmented views  $I_{k^n}^t$  and  $I_{k^n}^{t'}$ . We concatenate the negatives with queries and positives and build the batch of paired embeddings as  $q' = [q_1 \ q_2 \ \cdots \ q_N \ k_{1,t}^n \ k_{2,t}^n \ \cdots \ k_{N,t}^n]$  and  $k^{p'} = [k_1^p \ k_2^p \ \cdots \ k_N^p \ k_{1,t'}^n \ k_{2,t'}^n \ \cdots \ k_{N,t'}^n]$ .

At the inference stage, the preprocessed query and database images input the feature extractor. The preprocessing is detailed in Sec. G of the appendix.  $\tilde{q}$  and  $\tilde{k}$  denote the feature embeddings of  $I_q$  and  $I_k$ , respectively. By computing  $L_2$  distance between query and database feature embeddings, we use the K-Nearest Neighbors (KNN) algorithm to identify the top-K candidates.

**GeoPair:** The key idea of the GeoPair strategy is to merge metric learning from VG with the SSL paradigm. Metric learning leverages Hard Negative Mining (HNM) to form triplets or pairs that distinguish between positive ( $I_{k^p}$ ), negative ( $I_{k^n}$ ), and query samples ( $I_q$ ). SSL, on the other hand, uses data augmentation to create varied views ( $I_{k^n}^t$  and  $I_{k^n}^{t'}$ ) of a single instance as positives, treats other instances as negatives, and employs advanced loss functions to enhance sampling efficiency. GeoPair builds geo-specific query-positive pairs using geo-labels and incorporates augmented views of negatives, facilitating the learning of real geographic relationships and robust representation learning following the SSL paradigm.

#### 3.1. Loss functions

The VG-SSL framework integrates selected SSL loss functions, highlighting a critical distinction between metric learning and SSL in terms of the embeddings targeted by the loss function. Specifically, metric learning directly optimizes the distance metrics of feature embeddings produced by  $F_g$ . On the other hand, SSL approaches optimize projected embeddings generated by  $F_p$  to be invariant to data augmentation. Subsequently, during the inference phase, SSL methods revert to utilizing the original, unprojected feature embeddings output by  $F_g$ .

**InfoNCE Loss.** InfoNCE Loss [35] is used in contrastive learning SSL methods [13, 14, 26]. The intuition of this loss is to reduce the embedding distance between positive pairs and increase the embedding distance to all other pairs. The loss function is formulated as

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{N} \sum_{b=1}^N \log \frac{\exp(s(q_b, k_b^p)/\tau)}{\sum_{i=1, i \neq b}^N \exp(s(q_b, k_i^p)/\tau)}, \quad (1)$$

where  $\tau$  is the temperature parameter, and  $s(q_b, k_b^p) = \frac{q_b \cdot k_b^p}{\|q_b\|_2 \|k_b^p\|_2}$ .

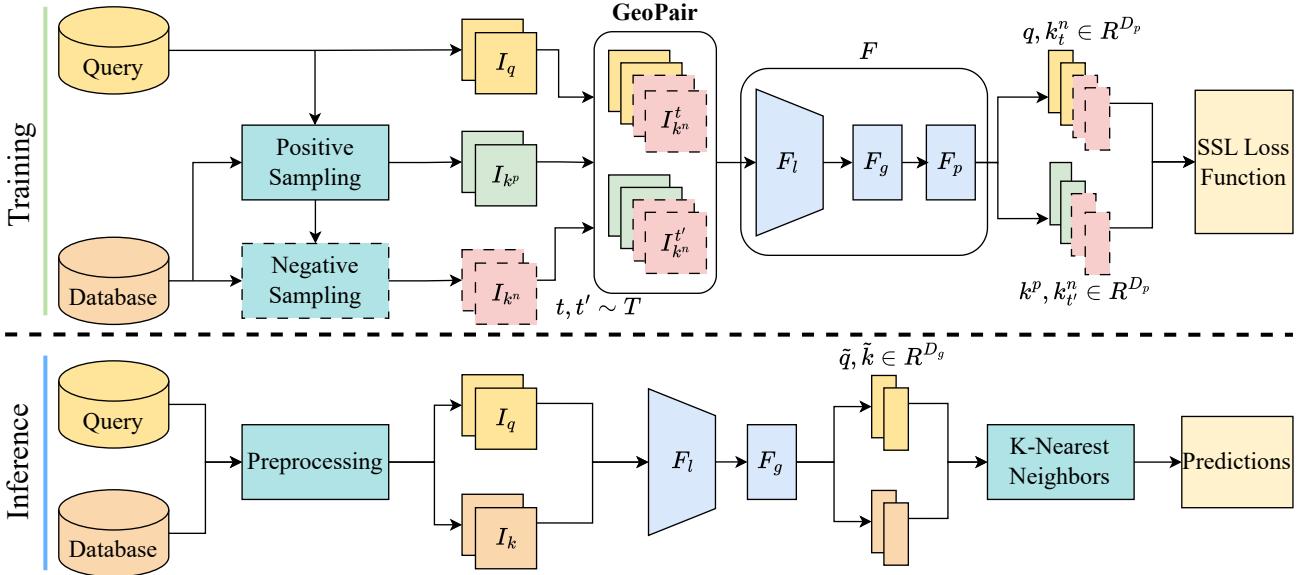


Figure 2. VG-SSL Architecture: During training, query images  $I_q$  and positive database images  $I_{k^p}$  are sampled, with optional negative images  $I_{k^n}$  selected via HNM. **GeoPair** strategy builds image pairs using query-positive pairs  $I_q, I_{k^p}$  and augmented negative pairs  $I_{k^n}^t, I_{k'^n}^{t'}$  with augmentation  $t, t' \sim T$ . The feature extractor  $F$  then produces embeddings ( $q, k^p, k_t^n$ , and  $k_{t'}^n$ ), and SSL loss is applied to train  $F$ . During inference, the projection head is removed, and KNN is used with feature embeddings  $\tilde{q}$  and  $\tilde{k}$  from the aggregation module.

**Embedding Prediction Loss.** Embedding Prediction Loss is used in self-distillation methods [15, 23]. The loss function aims to predict the common representation between the teacher embedding  $k_b^p$  and the student embedding  $q_b$ . The loss function is formulated as

$$\mathcal{L}_{\text{EmbPred}} = \frac{1}{N} \sum_{b=1}^N [2 - 2 \frac{p(q_b) \cdot \text{stopgrad}(k_b^p)}{\|p(q_b)\|_2 \| \text{stopgrad}(k_b^p) \|_2}], \quad (2)$$

where  $p(\cdot)$  is a shallow MLP to predict  $k_b^p$  from  $q_b$  and  $\text{stopgrad}(\cdot)$  means the gradient is not backpropagated to the teacher model.

**Cross-correlation Loss.** Barlow Twins (BT) [54] introduces the Cross-correlation Loss. First, it builds a cross-correlation matrix defined as

$$C_{ij} = \frac{\sum_{b=1}^N q_{b,i} k_{b,j}^p}{\sqrt{\sum_{b=1}^N (q_{b,i})^2} \sqrt{\sum_{b=1}^N (k_{b,j}^p)^2}}, \quad (3)$$

where  $i, j$  are the index of elements of the embedding, and  $C$  is a square cross-correlation matrix range from -1 (negative correlation) to 1 (positive correlation). Then, the BT loss  $L_{\text{BT}}$  is calculated as

$$\mathcal{L}_{\text{BT}} = \sum_{i=1}^D (1 - C_{ii})^2 + \lambda \sum_{i=1}^D \sum_{j=1, j \neq i}^D C_{ij}, \quad (4)$$

where  $\lambda$  is the weight of the off-diagonal term as a hyperparameter. The intuition of this loss is to enforce the same

element of embeddings positively correlated and the different elements of embeddings uncorrelated to maximize contained information.

**VIC Regularization Loss.** VIC Regularization (VICReg) Loss [6] is inspired by BT Loss and explicitly defines the invariance (first row), variance (second row), and covariance (third row) terms of the elements of the embeddings as

$$\begin{aligned} \mathcal{L}_{\text{VICReg}} = & \frac{\lambda_1}{ND} \sum_{b=1}^N \sum_{i=1}^D \|q_{b,i} - k_{b,i}^p\|_2^2 \\ & + \frac{\lambda_2}{D} \sum_{i=1}^D [\max\{d_2 - \text{std}_i(q), 0\} + \max\{d_2 - \text{std}_i(k^p), 0\}] \\ & + \frac{\lambda_3}{D} \sum_{i=1}^D \sum_{j=1, j \neq i}^D [\text{cov}_{ij}^2(q) + \text{cov}_{ij}^2(k^p)], \end{aligned} \quad (5)$$

where  $\lambda_1, \lambda_2, \lambda_3, d_2$  are scalars,  $\text{std}_i(q)$  is the standard deviation of the  $i$ -th element of  $q$  over the batch, and  $\text{cov}_{ij}(q)$  is the element of the  $i$ -th row and  $j$ -th column of the covariance matrix of  $q$  over the batch. This loss function decorrelates elements of the embedding by reducing the off-diagonal elements of the covariance matrix to zeros. It also maximizes the standard deviation of the element of the embeddings over the batch with a margin  $d_2$ .

## 4. Experimental Setup

### 4.1. Datasets and Metrics

We employ five VG datasets featuring handheld and car-mounted camera settings that are widely used in robotics and autonomous driving applications: Pitts30k [43], MSLS [50], Tokyo24/7 [42], Nordland [40], and Revisited San Francisco (R-SF) [12, 44] (Reported in Appendix Sec. B). We leverage the expansive MSLS dataset for both training and evaluation. For evaluation, the Pitts30k and Tokyo24/7 datasets are primarily composed of urban imagery, with both their database images derived from Google Street View. The query set of Tokyo24/7 is distinct, consisting of images captured via smartphone. The Nordland dataset presents a similar viewpoint to MSLS and R-SF but specifically introduces challenges related to seasonal change from summer to winter.

Our experimental evaluation is anchored by the Recall@N (R@N) metric, which assesses the proportion of query images that are correctly identified within the top-N predictions, provided that they fall within a pre-set proximity of 25 meters from the true geo-location. This threshold aligns with prior research in the VG domain for a fair comparison. This metric is a pivotal indicator of VG accuracy, with our analysis focusing on R@1, R@5, and R@10.

### 4.2. Implementation Details

In our study, we adopt certain protocols from the benchmark VG pipeline as outlined in [9], while incorporating GeoPair to accommodate various SSL techniques within the VG framework. The training image size is  $480 \times 640$  ( $H \times W$ ). With a batch size  $B$  of 64, each batch consists of 64 image pairs. We use Adam optimizer [29] with learning rates  $1e^{-5}$ . We instantiate both one-stage and two-stage methods with different SSL methods. For one-stage methods, we conduct our experiments mainly with ResNet-50 [27] + GeM [36] or DeiT-S [45] (Distilled version of ViT-S [19]) used in [9, 56]. For two-stage methods, we primarily use DeiT-S for the first stage and the reranking part of R2Former [56] for the second stage. We run each experiment for 36 hours with approximately 200 epochs. The data augmentation is set to be random resizing and cropping with a random scale from 1.0 to 1.25, and random flipping ( $p = 0.5$ ). During training, we define images as positive samples if they are within the 10 m radius of the query images and negative samples if they are beyond the 25 m radius of the query images. The models are primarily trained with a batch size of 64 (ResNet-50 + GeM) or 32 (DeiT-S) in one NVIDIA-A100-80GB GPU, while the final model is trained with a batch size of 32, 64, 128, or 256, depending on the final performance. The pipelines are implemented using Pytorch-lightning [20].

**Remarks on Data Augmentation.** As noted in [9], the

effectiveness of data augmentation can vary across datasets due to their unique characteristics. To establish a robust baseline, we fully utilize standard data augmentations, including random flipping and random resized cropping, which avoid dataset-specific biases and provide general benefits across all datasets. While we agree that exploring more advanced augmentations could enhance performance, applying them across multiple datasets can be complex and may result in inconsistent outcomes. We suggest investigating the impact of specific augmentations when fine-tuning models on individual datasets for more targeted improvements.

## 5. Results

### 5.1. Comparison with State-of-the-art

Table 1 presents a comparison between the best results of various SSL methods and state-of-the-art one-stage VG methods, including NetVLAD [4], SFRS [22], GCL-GeM [31], TransVPR [47] (without re-ranking), and R2Former [56] (without re-ranking) for both ResNet-50 [27] and ResNeXt [52]. To ensure a fair comparison, all methods were trained on the MSLS and Pitts30k datasets. Details about the state-of-the-art results can be found in Sec. F of the appendix. Due to page limitations, extended results for two-stage methods, including SP-SuperGlue [38], Patch-NetVLAD [25], TransVPR [47], and R2Former [56], are provided in Sec. E of the appendix.

**Inference Complexity:** For our one-stage methods, we use ResNet-50 [27] + GeM [36] with  $D_g = 1024$  and DeiT-S [45] with  $D_g = 256$ . This ensures that memory usage and inference time are consistent with ResNet50-GeM [9] and R2Former [56] without re-ranking.

In Table 1, our comparison of different Self-Supervised Learning (SSL) strategies shows that SimCLR, MoCov2, and Barlow Twins consistently perform well across various datasets, including MSLS (Val and Challenge subsets), Pitts30k, and Nordland, demonstrating their versatility and effectiveness in diverse environments, from general landscapes to urban settings. SimCLR stands out for its consistently strong performance across all datasets, highlighting the robustness of contrastive learning. These results emphasize the power of both contrastive learning and information maximization strategies.

When comparing ResNet50-GeM ( $D_g = 1024$ ) and DeiT-S ( $D_g = 256$ ) variants, we observed a drop in performance for SimCLR, MoCov2, BYOL, and SimSiam on MSLS and Nordland, particularly affecting the R@1 metric in contrastive learning methods and overall recall in self-distillation methods. This indicates the significant impact of embedding dimension ( $D_g$ ) on performance. In contrast, information maximization methods like Barlow Twins and VICReg maintained or improved their metrics with a

Table 1. Comparison of state-of-the-art VG methods with our results on large-scale VG datasets. Our models were initially trained on the MSLS dataset. For evaluating performance in urban environments (Pitts30k and Tokyo24/7), we further fine-tuned our models on the Pitts30k dataset. The best results are highlighted in **bold**, with colors indicating performance tiers of our results: green for the top 3 results, yellow for the next 2 results, and red for the 2 lowest-performing results. \* denotes the performance without reranking.

$D_g$	MSLS Val			MSLS Challenge			Pitts30k			Tokyo24/7			Nordland			
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	
<i>One-Stage Methods</i>																
NetVLAD [4]	-	60.8	74.3	79.5	35.1	47.4	51.7	81.9	91.2	93.7	64.8	78.4	81.6	-	-	
SFRS [22]	4096	69.2	80.3	83.1	41.5	52.0	56.3	<b>89.4</b>	<b>94.7</b>	95.9	<b>85.4</b>	<b>91.1</b>	<b>93.3</b>	18.8	32.8	39.8
TransVPR* [47]	256	70.8	85.1	89.6	48.0	67.1	73.6	73.8	88.1	91.9	-	-	-	15.9	38.6	49.4
R2Former* [56]	256	79.3	90.5	92.7	54.9	75.1	79.6	72.9	88.5	92.6	43.5	65.7	72.4	21.4	33.7	41.0
GCL-ResNet50-GeM [31]	1024	74.6	84.7	88.1	52.9	65.7	71.9	79.9	90.0	92.8	58.7	71.1	76.8	-	-	-
GCL-ResNeXt-GeM [31]	1024	80.9	90.7	92.6	62.3	76.2	81.1	79.2	90.4	93.2	58.1	74.3	78.1	-	-	-
<i>Our One-Stage Methods with ResNet50-GeM</i>																
Triplet Loss (Baseline) [8]	1024	<b>76.9</b>	<b>86.1</b>	<b>89.5</b>	<b>53.5</b>	<b>68.1</b>	<b>72.3</b>	<b>76.7</b>	<b>89.1</b>	<b>92.3</b>	<b>50.2</b>	<b>67.9</b>	<b>76.8</b>	<b>39.5</b>	<b>59.0</b>	<b>67.7</b>
SimCLR	1024	<b>84.2</b>	<b>92.2</b>	<b>94.2</b>	<b>63.1</b>	<b>78.9</b>	<b>83.6</b>	82.8	91.9	94.6	54.6	74.9	81.9	<b>39.9</b>	56.4	63.9
MoCov2	1024	81.5	90.5	92.8	59.0	73.8	79.2	82.6	92.4	95.1	51.4	68.3	76.5	28.0	42.7	50.1
BYOL	1024	<b>72.7</b>	<b>85.5</b>	<b>87.7</b>	<b>50.4</b>	<b>66.4</b>	<b>71.4</b>	80.2	91.5	94.4	<b>44.8</b>	<b>63.8</b>	<b>70.8</b>	10.6	18.5	23.5
SimSiam	1024	<b>75.0</b>	<b>85.8</b>	<b>88.6</b>	<b>52.1</b>	<b>67.0</b>	<b>72.2</b>	<b>78.6</b>	<b>89.8</b>	<b>92.7</b>	51.1	67.6	71.4	12.5	21.5	27.0
Barlow Twins	1024	<b>79.5</b>	<b>89.5</b>	<b>91.9</b>	<b>59.2</b>	<b>74.2</b>	<b>79.1</b>	80.8	91.7	94.2	<b>45.7</b>	<b>61.9</b>	<b>70.8</b>	18.5	30.5	38.0
VICReg	1024	77.4	89.3	91.2	58.0	74.1	79.0	80.2	91.3	94.1	50.2	65.4	74.3	14.9	25.1	31.3
<i>Our One-Stage Methods with DeiT-S</i>																
Triplet Loss (Baseline) [56]	256	<b>79.3</b>	<b>90.5</b>	<b>92.7</b>	<b>54.9</b>	<b>75.1</b>	<b>79.6</b>	<b>72.9</b>	<b>88.5</b>	<b>92.6</b>	<b>43.5</b>	<b>65.7</b>	<b>72.4</b>	21.4	33.7	41.0
SimCLR	256	81.1	91.1	93.1	58.9	77.1	82.6	84.7	93.9	<b>96.0</b>	59.4	76.2	80.0	24.9	38.9	46.1
MoCov2	256	<b>76.1</b>	<b>88.5</b>	<b>91.1</b>	<b>56.8</b>	<b>75.2</b>	<b>78.7</b>	80.8	92.4	95.0	50.8	69.8	77.1	15.4	26.4	33.0
BYOL	256	<b>58.2</b>	<b>75.3</b>	<b>79.6</b>	<b>37.7</b>	<b>54.0</b>	<b>60.4</b>	<b>76.6</b>	<b>89.4</b>	<b>92.9</b>	43.2	62.2	68.6	4.1	7.9	10.6
SimSiam	256	<b>56.2</b>	<b>76.2</b>	<b>80.1</b>	<b>35.3</b>	<b>52.3</b>	<b>58.7</b>	79.7	91.0	93.6	47.3	<b>63.8</b>	<b>74.0</b>	6.2	11.5	15.4
Barlow Twins	256	<b>79.7</b>	<b>91.4</b>	<b>93.1</b>	<b>59.1</b>	<b>76.1</b>	<b>81.5</b>	82.6	92.1	95.0	58.4	<b>75.2</b>	<b>80.6</b>	28.1	43.3	51.1
VICReg	256	75.8	89.5	91.9	56.9	74.0	78.2	81.7	92.3	95.2	51.7	66.7	74.6	19.3	32.1	39.6

smaller  $D_g$ , suggesting greater robustness to reduced dimensions.

Compared to state-of-the-art one-stage methods, SimCLR, MoCov2, and Barlow Twins exhibit competitive or superior performance on the MSLS and Nordland datasets, likely due to their advanced feature learning capabilities. SFRS [22], while excelling in urban datasets like Pitts30k and Tokyo24/7, relies on a much larger embedding dimension ( $D_g = 4096$ ), highlighting a trade-off between computational complexity and performance. However, SFRS shows limited generalization in more diverse environments like MSLS and Nordland, possibly due to its training bias toward urban landscapes.

On the other hand, GCL methods [31] perform similarly to our approaches with  $D_g = 1024$ , but they require additional processing steps like Principal Component Analysis (PCA) and orientation label training. In contrast, our SSL-based methods avoid such extra steps, using smaller embedding dimensions ( $D_g = 1024$  for ResNet50-GeM and  $D_g = 256$  for DeiT-S), reducing computational overhead and improving inference efficiency. This underscores not only the competitiveness of our SSL strategies with state-of-the-art methods but also the practical efficiency of our approach.

## 5.2. Ablation study

In this section, we conduct an ablation study with our one-stage methods on the following key factors influencing VG performance: Hard Negative Mining (HNM), Number of Projection Layers  $L$ , Dimensionality of Projected Embeddings  $D_p$  and Feature Embeddings  $D_g$ . We report

MSLS Val's performance for the study. If not specifically mentioned, we set the following parameters from the initial hyperparameter search:  $F$  is ResNet50-GeM; Training batch size is 64; SimCLR, BT, and VICReg use HNM; MoCov2, BYOL, and SimSiam do not use HNM;  $D_p = 2048$ ;  $D_g = 1024$ ; SimCLR, MoCov2 use  $L = 1$ ; BYOL, SimSiam, BT, VICReg use  $L = 2$ .

### 5.2.1 Hard Negative Mining

In Table 2, we present the results associated with training models with and without the implementation of HNM. It is evident that models such as SimCLR, BT, and VICReg demonstrate enhanced performance when HNM is applied. Conversely, MoCov2 and SimSiam exhibit superior performance in the absence of HNM. The results for BYOL exhibit variability in performance with and without HNM application. However, the notably higher R@1 without HNM suggests a slight preference for training without HNM.

### 5.2.2 Number of Projection Layers

In Table 2, we analyze the impact of the number of projection layers ( $L$ ) on SSL model convergence. Notably, methods like BYOL, SimSiam, and VICReg fail to converge with fewer than two layers, as a single layer only supports linear transformations, while two or more layers enable more complex, non-linear transformations that enhance the representational power of the embeddings. In contrast, SimCLR performs best with one layer, while MoCov2 excels with none. Both BYOL and SimSiam perform well at  $L = 2$ ,

Table 2. Comparison of training with hard negative mining, different number of projection layers  $L$ , different dimension of projected embeddings  $D_p$ , and different dimension of feature embeddings  $D_g$ . The best results for each setting are highlighted in **bold**.  $\times$  means that the model cannot converge. - means the model is skipped since the performance has saturated.

SimCLR			MoCov2			BYOL			SimSiam			Barlow Twins			VICReg			
R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	
<i>Hard Negative Mining</i>																		
79.6	90.8	92.0	<b>77.8</b>	<b>89.3</b>	<b>91.9</b>	<b>73.0</b>	83.2	85.9	<b>74.3</b>	<b>84.9</b>	<b>87.4</b>	65.5	82.7	85.8	66.1	81.4	85.3	
✓	<b>82.4</b>	<b>91.4</b>	<b>93.4</b>	76.5	86.4	88.2	71.1	<b>83.9</b>	<b>86.2</b>	71.8	83.0	85.7	<b>79.5</b>	<b>89.5</b>	<b>91.9</b>	<b>77.6</b>	<b>86.9</b>	<b>89.2</b>
<i>Number of Projection Layers L</i>																		
0	80.9	<b>91.9</b>	93.1	<b>80.5</b>	<b>90.1</b>	<b>91.9</b>	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	76.6	86.6	88.5	$\times$	$\times$	$\times$
1	<b>82.4</b>	91.4	<b>93.4</b>	77.8	89.3	<b>91.9</b>	$\times$	$\times$	$\times$	$\times$	$\times$	$\times$	78.6	87.6	89.5	$\times$	$\times$	$\times$
2	<b>82.4</b>	90.1	91.8	77.0	88.0	89.6	<b>73.0</b>	83.2	<b>85.9</b>	<b>74.3</b>	<b>84.9</b>	<b>87.4</b>	<b>79.5</b>	<b>89.5</b>	<b>91.9</b>	<b>77.6</b>	86.9	89.2
3	-	-	-	-	-	-	70.4	<b>83.8</b>	85.5	69.3	83.9	87.3	78.8	89.1	91.4	76.9	<b>87.8</b>	<b>90.0</b>
<i>Dimension of projected embeddings <math>D_p</math></i>																		
1024	<b>84.2</b>	<b>92.2</b>	<b>94.3</b>	78.5	89.1	<b>93.2</b>	71.5	83.9	86.9	72.4	84.5	87.0	76.4	87.6	89.9	74.1	86.1	88.2
2048	82.4	91.4	93.4	77.8	<b>89.3</b>	91.9	<b>73.0</b>	83.2	85.9	74.3	84.9	87.4	<b>79.5</b>	<b>89.5</b>	<b>91.9</b>	77.6	86.9	89.2
4096	82.3	91.8	93.0	<b>79.9</b>	<b>89.3</b>	91.4	72.8	<b>85.5</b>	<b>87.7</b>	<b>75.0</b>	<b>85.8</b>	<b>88.6</b>	76.1	87.7	90.4	<b>77.7</b>	<b>88.8</b>	<b>90.7</b>
<i>Dimension of feature embeddings <math>D_g</math></i>																		
512	81.6	<b>91.5</b>	<b>93.8</b>	<b>79.5</b>	89.2	90.8	71.9	84.9	87.6	72.8	86.4	88.6	76.4	87.8	90.7	77.3	<b>87.7</b>	89.7
1024	82.4	91.4	93.4	77.8	<b>89.3</b>	<b>91.9</b>	73.0	83.2	85.9	74.3	84.9	87.4	<b>79.5</b>	<b>89.5</b>	<b>91.9</b>	77.6	86.9	89.2
2048	<b>83.0</b>	91.2	92.6	78.4	89.1	91.4	<b>74.6</b>	<b>85.8</b>	<b>88.4</b>	<b>74.6</b>	<b>86.4</b>	<b>88.8</b>	79.2	89.3	91.4	<b>78.5</b>	<b>87.7</b>	<b>89.9</b>

and BT shows improved performance with two layers. VICReg’s performance varies between  $L = 2$  and  $L = 3$ , indicating a more complex relationship between projection layers and VG performance.

### 5.2.3 Projected Embedding Dimensionality

In examining the effect of projected embedding dimensions ( $D_p$ ) on retrieval performance (Table 2), we find a non-linear relationship between embedding size and metrics (R@1, R@5, R@10). SimCLR performs best at the lowest dimension (1024), with R@1 at 84.2%, R@5 at 92.2%, and R@10 at 94.3%, while increasing  $D_p$  to 2048 and 4096 slightly reduces performance.

In contrast, MoCov2, BYOL, SimSiam, and VICReg generally improve with larger  $D_p$ , with VICReg reaching its peak at 4096 (R@1: 77.7%, R@5: 88.8%, R@10: 90.7%). BT, however, performs best at  $D_p$  of 2048. This indicates that the optimal embedding dimension depends on the method and the interaction between model architecture and the complexity of the projected space.

### 5.2.4 Feature Embeddings Dimensionality

In Table 2, we present the results for different feature embedding dimensions ( $D_g$ ). The embedding dimension is key for memory consumption, inference and matching time in VG architectures, and it plays a critical role in improving model performance and representation ability.

Our data shows that SimCLR benefits from a larger  $D_g$  (2048), with a notable increase in R@1, though R@5 and R@10 remain stable across dimensions. In contrast, MoCov2 performs best at lower  $D_g$  values (512 and 1024), indicating less sensitivity to larger dimensions. BYOL, SimSiam, and VICReg favor the highest  $D_g$  (2048), while BT’s performance plateaus at  $D_g = 1024$ , showing diminishing returns beyond this point.

## 5.3. Visualization

In this study, we analyze activation maps from the final layers of ResNet-50 models trained with Triplet Loss (Baseline) and SSL methods, as shown in Fig.3. Our findings reveal that, across all three datasets, models trained with Triplet Loss exhibit diffuse activation over irrelevant features, with inconsistent patterns under day-to-night shifts (MSLS) and viewpoint changes (Pitts30k). In the Tokyo24/7 dataset, these models mistakenly highlight dynamic objects like vehicles and roads. In contrast, our models, trained with SSL paradigm, focus more consistently on landmark regions and show uniform activation patterns. We also review the top-5 retrieval results of SSL methods on the MSLS validation set in Sec.H of the appendix, demonstrating their effectiveness.

## 5.4. Discussion

The optimal training settings for various SSL methods (Table 3) underscore the importance of method-specific configurations to maximize performance. Our study highlights Hard Negative Mining (HNM), the number of projection layers, and the dimensions of projected and feature embeddings as key factors influencing geo-specific representations. We discuss HNM and projector settings by category and compare optimal SSL settings between general applications and VG, offering guidance for enhancing SSL in geo-related tasks.

**Contrastive Learning.** HNM improves SimCLR’s performance but not MoCov2’s, likely due to MoCov2’s memory dictionary, which reduces the need for HNM. HNM may also reduce the diversity of MoCov2’s memory bank. Additionally, multiple projection layers are less critical for both SimCLR and MoCov2 in VG than in general applications.

**Self-distillation Learning.** These methods show less reliance on HNM, likely due to their loss functions. Optimal performance is achieved with two projection layers and

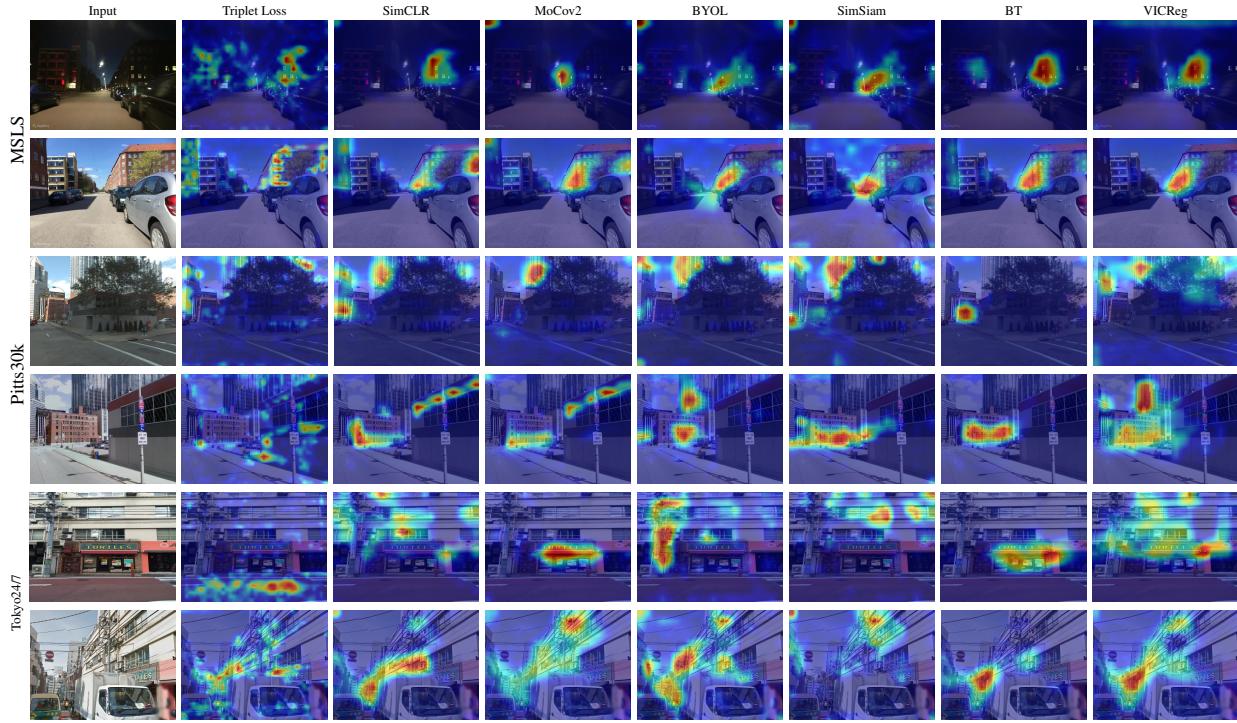


Figure 3. The activation maps of ResNet-50 models trained with Triplet Loss (Baseline) and SSL methods. For each dataset, the first row is for the query image and the second row is for the positive sample image.

Table 3. Optimal training settings for SSL methods in VG and general applications from original papers. For general applications, HNM is not applied,  $D_{ph}$  is the hidden layer dimension of the projector, and  $D_{po}$  is the output dimension of the projector.

Categories	Methods	VG			General		
		HNM	L	$D_p$	L	$D_{ph}$	$D_{po}$
Contrastive Learning	SimCLR	✓	1	1024	2	2048	128
	MoCov2	0	-	-	2	2048	128
Self-distillation Learning	BYOL		2	4096	2	4096	256
	SimSiam		2	4096	3	2048	2048
Information Maximization	Barlow Twins	✓	2	2048	3	8192	8192
	VICReg	✓	3	4096	3	8192	8192

larger embedding dimensions, suggesting a preference for complex, high-dimensional embeddings. Notably, configurations differ between VG and general applications, with BYOL and SimSiam favoring different projector setups in VG. For example, BYOL prefers larger projected embeddings in VG compared to general application scenarios and SimSiam prefers a 2-layer projector instead of a 3-layer one.

**Information Maximization.** HNM benefits methods like Barlow Twins, which performs best with a 2-layer projection head and mid-range embedding sizes in VG, differing from its general application setup. VICReg achieves optimal results with three projection layers and large embedding dimensions, consistent with its general preferences.

## 6. Conclusions

In this paper, we introduced the unified VG-SSL framework, which benchmarks various SSL methods alongside the GeoPair strategy for learning geo-specific representations in VG tasks. We evaluated six prominent SSL approaches for VG tasks in robotics and autonomous vehicles datasets, identifying optimal SSL settings, with a strong preference for contrastive learning and information maximization methods. SimCLR, MoCov2, and Barlow Twins emerged as top performers, matching or exceeding the performance of leading VG methods. Our findings emphasize the potential of SSL for VG tasks, and the flexibility of our framework enables its application to benchmark geo-specific representation learning problems. Additionally, the adaptable GeoPair strategy offers an opportunity extended to other retrieval-based tasks, such as face recognition. Future work will focus on integrating more SSL methods and exploring additional augmentations.

**Acknowledgement.** This work was supported by the Technology Innovation Institute, the NSF CAREER Award 2145277, the NSF CPS Grant CNS-2121391, and the NYU IT High Performance Computing resources, services, and staff expertise. Giuseppe Loianno serves as consultant for the Technology Innovation Institute. This arrangement has been reviewed and approved by the New York University in accordance with its policy on objectivity in research.

## References

- [1] Ali-bey, A., Chaib-draa, B., Giguère, P.: Global proxy-based hard mining for visual place recognition. arXiv preprint arXiv:2302.14217 (2023) [3](#)
- [2] Ali-bey, A., Chaib-draa, B., Giguère, P.: Mixvpr: Feature mixing for visual place recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2998–3007 (2023) [1](#), [2](#)
- [3] Ali-bey, A., Chaib-draa, B., Giguère, P.: Gsv-cities: Toward appropriate supervised visual place recognition. Neurocomputing **513**, 194–203 (2022) [2](#)
- [4] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307 (2016) [1](#), [2](#), [5](#), [6](#), [13](#), [14](#)
- [5] Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., Ermon, S.: Geography-aware self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10181–10190 (2021) [3](#)
- [6] Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In: ICLR (2022) [2](#), [3](#), [4](#), [13](#)
- [7] Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013) [1](#)
- [8] Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4878–4888 (June 2022) [1](#), [2](#), [6](#)
- [9] Berton, G., Mereu, R., Trivigno, G., Masone, C., Csurka, G., Sattler, T., Caputo, B.: Deep visual geo-localization benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5396–5407 (2022) [1](#), [2](#), [5](#), [12](#), [13](#)
- [10] Berton, G., Trivigno, G., Caputo, B., Masone, C.: Eigenplaces: Training viewpoint robust models for visual place recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11080–11090 (October 2023) [2](#)
- [11] Chen, C., Liu, X., Xu, X., Li, Y., Ding, L., Wang, R., Feng, C.: Self-supervised visual place recognition by mining temporal and feature neighborhoods. arXiv preprint arXiv:2208.09315 (2022) [1](#)
- [12] Chen, D.M., Baatz, G., Köser, K., Tsai, S.S., Vedantham, R., Pylvänen, T., Roimela, K., Chen, X., Bach, J., Pollefeys, M., Girod, B., Grzeszczuk, R.: City-scale landmark identification on mobile devices. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 737–744 (2011) [5](#), [12](#)
- [13] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (13–18 Jul 2020) [2](#), [3](#), [13](#)
- [14] Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) [2](#), [3](#), [13](#)
- [15] Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15750–15758 (June 2021) [2](#), [3](#), [4](#), [12](#), [13](#)
- [16] Chen, Z., Jacobson, A., Sünderhauf, N., Upcroft, B., Liu, L., Shen, C., Reid, I., Milford, M.: Deep learning features at scale for visual place recognition. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 3223–3230 (2017) [1](#)
- [17] Costante, G., Ciarfuglia, T.A., Valigi, P., Ricci, E.: A transfer learning approach for multi-cue semantic place recognition. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 2122–2129 (2013) [2](#)
- [18] Cummins, M., Newman, P.: Fab-map: Probabilistic localization and mapping in the space of appearance. The International Journal of Robotics Research **27**(6), 647–665 (2008) [2](#)
- [19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021) [5](#)
- [20] Falcon, W., The PyTorch Lightning team: PyTorch Lightning (Mar 2019), <https://github.com/Lightning-AI/lightning> [5](#)

- [21] Garg, S., Suenderhauf, N., Milford, M.: Semantic–geometric visual place recognition: a new perspective for reconciling opposing views. *The International Journal of Robotics Research* **41**(6), 573–598 (2022) 1
- [22] Ge, Y., Wang, H., Zhu, F., Zhao, R., Li, H.: Self-supervising fine-grained region similarities for large-scale image localization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 369–386. Springer (2020) 2, 5, 6, 14
- [23] 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: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS’20, Curran Associates Inc., Red Hook, NY, USA (2020) 2, 4, 13
- [24] Gálvez-López, D., Tardós, J.D.: Real-time loop detection with bags of binary words. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 51–58 (2011) 2
- [25] Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: 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). pp. 14141–14152 (June 2021) 2, 5, 12, 13, 14
- [26] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 2, 3
- [27] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 3, 5
- [28] Khaliq, A., Milford, M., Garg, S.: Multires-netvlad: Augmenting place recognition training with low-resolution imagery. *IEEE Robotics and Automation Letters* **7**(2), 3882–3889 (2022) 2
- [29] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 5
- [30] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *nature* **521**(7553), 436–444 (2015) 1, 2
- [31] Leyva-Vallina, M., Strisciuglio, N., Petkov, N.: Data-efficient large scale place recognition with graded similarity supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23487–23496 (June 2023) 2, 5, 6, 13, 14
- [32] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International journal of computer vision* **60**(2), 91–110 (2004) 1
- [33] Milford, M.J., Wyeth, G.F.: Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In: 2012 IEEE International Conference on Robotics and Automation. pp. 1643–1649 (2012) 2
- [34] Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6707–6717 (2020) 1
- [35] Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 2, 3
- [36] Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(7), 1655–1668 (2019) 2, 3, 5
- [37] Razavian, A.S., Sullivan, J., Carlsson, S., Maki, A.: Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications* **4**(3), 251–258 (2016) 2
- [38] Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020) 2, 5, 12, 14
- [39] Saviolo, A., Loianno, G.: Learning quadrotor dynamics for precise, safe, and agile flight control. *Annual Reviews in Control* (2023) 1
- [40] Sünderhauf, N., Neubert, P., Protzel, P.: 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). p. 2013 (2013) 5
- [41] Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. arXiv preprint arXiv:1511.05879 (2015) 2
- [42] Torii, A., Arandjelović, R., Sivic, J., Okutomi, M., Pandla, T.: 24/7 place recognition by view synthesis. In:

- 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1808–1817 (2015) 5
- [43] Torii, A., Sivic, J., Pajdla, T., Okutomi, M.: Visual place recognition with repetitive structures. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 883–890 (2013) 5
- [44] Torii, A., Taira, H., Sivic, J., Pollefeys, M., Okutomi, M., Pajdla, T., Sattler, T.: Are large-scale 3d models really necessary for accurate visual localization? *IEEE transactions on pattern analysis and machine intelligence* **43**(3), 814–829 (2019) 5, 12
- [45] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021) 5
- [46] Trivigno, G., Berton, G., Aragon, J., Caputo, B., Massone, C.: Divide&classify: Fine-grained classification for city-wide visual geo-localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11142–11152 (October 2023) 2
- [47] Wang, R., Shen, Y., Zuo, W., Zhou, S., Zheng, N.: Transvpr: Transformer-based place recognition with multi-level attention aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13648–13657 (June 2022) 2, 5, 6, 12, 13, 14
- [48] Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5022–5030 (2019) 2
- [49] Wang, Y., Albrecht, C., Braham, N., Mou, L., Zhu, X.: Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine* **PP**, 2–36 (12 2022) 3
- [50] Warburg, F., Hauberg, S., López-Antequera, M., Gargallo, P., Kuang, Y., Civera, J.: Mapillary street-level sequences: A dataset for lifelong place recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2623–2632 (2020) 5
- [51] Xiao, J., Tortei, D., Roura, E., Loianno, G.: Long-range uav thermal geo-localization with satellite imagery. arXiv preprint arXiv:2306.02994 (2023) 1
- [52] Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017) 5
- [53] Yandex, A.B., Lempitsky, V.: Aggregating local deep features for image retrieval. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1269–1277 (2015) 2
- [54] Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 12310–12320. PMLR (18–24 Jul 2021) 2, 3, 4, 13
- [55] Zhang, X., Wang, L., Su, Y.: Visual place recognition: A survey from deep learning perspective. *Pattern Recognition* **113**, 107760 (2021) 2
- [56] Zhu, S., Yang, L., Chen, C., Shah, M., Shen, X., Wang, H.: R2former: Unified retrieval and reranking transformer for place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19370–19380 (June 2023) 2, 5, 6, 12, 13, 14

## A. Appendix Overview

This supplementary material provides the following additional information for a better understanding of our paper: Sec. B report our results on the Revisited San Francisco (R-SF) dataset. Sec. C discusses the limitations of our work. Sec. D summarizes the comparison of architecture for different SSL methods. Sec. E shows the extended results of our two-stage methods. Sec. F explains the data source of state-of-the-art VG results. Sec. G explains the preprocessing procedure in the inference stage. Sec H shows the analysis of visualized results of models trained with different SSL strategies.

## B. Results on R-SF

We add results on Revisited San Francisco (R-SF) [12, 44] in Table B, comparing with triplet loss baseline. The results reflect a similar trend as other datasets except that Barlow Twins does not show superior performance.

## C. Limitations

This section reflects on the limitations of our research. A key limitation is the small batch sizes used (32 and 64), which are significantly smaller than those typically used in SSL methods, often exceeding 512. This may result in suboptimal performance. We chose smaller batch sizes to maintain comparability with other VG studies due to resource constraints. Larger batch sizes would have required more training steps to achieve similar results.

Another limitation is the scope of data augmentation. As noted in the Deep VG benchmark [9], the effectiveness of data augmentation can vary across datasets due to their unique characteristics. To establish a robust baseline, we focused on fundamental augmentations, such as random flipping and random resized cropping, which avoid dataset-specific biases and provide general benefits across all datasets. While we agree that exploring more advanced augmentations could enhance performance, applying them across multiple datasets can be complex and may result in inconsistent outcomes. We suggest investigating the impact of specific augmentations when fine-tuning models on individual datasets for more targeted improvements.

## D. Architecture difference of SSL methods

Table 5 presents a comparative analysis of various SSL methods, focusing specifically on their architectural differences. SSL techniques within identical categories tend to exhibit similarities in their loss functions and overall architectural designs. This observation encourages an analysis of our results based on categorical groupings. Nevertheless, variations in SSL approaches within the same category may lead to discrepancies in outcomes and the selection of op-

Table 4. Comparison of SSL methods and Triplet Loss on R-SF dataset

	R-SF		
	R@1	R@5	R@10
<i>Our One-Stage Methods with ResNet50-GeM</i>			
Triplet Loss (Baseline)	44.6	58.0	63.2
SimCLR	46.8	64.0	69.1
MoCov2	40.5	55.2	59.7
BYOL	25.1	38.1	46.0
SimSiam	25.3	39.3	44.6
Barlow Twins	22.4	36.0	41.6
VICReg	22.2	36.5	43.0
<i>Our One-Stage Methods with DeiT-S</i>			
Triplet Loss (Baseline)	39.1	57.0	63.0
SimCLR	50.2	65.1	70.6
MoCov2	35.3	53.8	58.2
BYOL	17.6	29.3	36.0
SimSiam	15.7	28.8	33.6
Barlow Twins	36.5	52.8	58.0
VICReg	29.1	44.0	51.3

timal hyperparameters. For SimSiam [15], we specifically note that we remove the BN layer for the output of the projection head since it does not converge with that BN layer.

## E. Extended results of our two-stage methods

In Table 6, we show the extended results of the comparison between our two-stage methods with R2Former [56] for reranking and state-of-the-art two-stage methods (SP-SuperGlue [38], Patch-NetVLAD [25], TransVPR [47], and R2Former [56]). For better readability, we concatenate the one-stage results (Table 1) at the bottom of the table.

When analyzing different SSL training strategies, it becomes evident that two-stage methods, particularly those rooted in contrastive learning and information maximization, demonstrate superior performance compared to self-distillation approaches. This trend aligns with observations made in one-stage results, underscoring the inherited high geo-specific representation quality from the first-stage results. Notably, among these strategies, SimCLR and Barlow Twins stand out, delivering higher overall performance metrics than their counterparts.

In our comparative analysis of two-stage methods against leading approaches, we observed that SimCLR and Barlow Twins generally match or exceed the performance of existing state-of-the-art methods across most datasets, with the notable exception of the Tokyo24/7 dataset. Despite this, our enhancements to the original R2Former model yield only marginal gains, even though our initial-stage results surpass those of the original R2Former. To analyze the training bottleneck in two-stage methods, we conducted a detailed examination of the MSLS dataset’s validation performance across various training stages, as outlined in Table 7.

Table 5. Comparison of architecture for selected SSL methods. **ME**: Momentum target encoder. **SG**: Stop gradient for target encoder. **PR**: Predictor to infer target (teacher) embeddings based on online (student) embeddings. **BN**: Batch normalization in the projector or predictor. **LP**: Large dimensionality of projected embeddings.

Categories	Methods	ME	SG	PR	BN	LP	Loss Function
Contrastive Learning	SimCLR [13]						InfoNCE Loss
	MoCov2 [14]		✓				InfoNCE Loss
Self-distillation Learning	BYOL [23]	✓	✓	✓	✓		Embedding Prediction Loss
	SimSiam [15]		✓	✓	✓		Embedding Prediction Loss
Information Maximization	Barlow Twins [54]				✓	✓	Cross-correlation Loss
	VICReg [6]				✓	✓	VIC Regularization Loss

We pay particular attention to the R@1 metric. In the global-retrieval-training phase, the performance ranking is as follows: SimCLR > Barlow Twins > R2Former > MoCov2 > VICReg > BYOL > SimSiam. However, this order shifts in the reranking-training phase to: SimCLR > VICReg > R2Former > Barlow Twins > MoCov2 > SimSiam > BYOL. This reshuffling illustrates a complex relationship between the outcomes of the global-retrieval-training and reranking-training stages, indicating that superior performance in the former does not automatically translate to enhanced results in the latter.

Additionally, our findings reveal that post-finetuning, most variants reach a plateau in MSLS validation performance, with R@1 nearing 90%, R@5 around 95%, and R@10 close to 96%. This saturation suggests a limit to the efficacy of the current fine-tuning approaches, highlighting the need for better strategies to push these metrics further.

## F. Data Source of State-of-the-art VG Results

In presenting the state-of-the-art results in VG methods as detailed in Table 1 and Table 6, we primarily draw upon data from several key research papers: R2Former [56], TransVPR [47], Patch-NetVLAD [25], and GCL [31]. When considering the NetVLAD [4] result, it is pertinent to acknowledge the variety of results yielded by different reproductions. For the purposes of this study, we refer to the results as documented by R2Former [56]. However, it is notable that the results for the Nordland datasets and the dimension of the feature are not included in the paper, a gap attributed to the lack of clarity regarding the source of their reproduction. For the results of R2Former without reranking part, we download the model provided by the authors and evaluate it across VG datasets.

## G. Preprocessing for different datasets

In the inference stage, due to the potential variability in the input images, preprocessing becomes an essential step. Our one-stage methods, as outlined in Table 1, predominantly utilize resizing as a key preprocessing technique.

This ensures that the dimensions of the input images align with those used during training. However, an exception is noted for the Tokyo24/7 dataset. Here, standard resizing procedures would alter the aspect ratios of the query images, potentially degrading performance. To address this, we adopt the *single\_query* preprocessing approach, as described in [9], specifically for the preprocessing of query images.

## H. Visualization results

In Fig. 4 - 7, we present a comparative analysis of the top-5 retrieved images from the MSLS validation dataset, using ResNet50-GeM models trained via different self-supervised learning (SSL) strategies. This qualitative assessment specifically addresses challenges such as illumination change, seasonal change, viewpoint change, and occlusion. Our findings reveal that SimCLR, MoCov2, and Barlow Twins consistently outperform other methods in tackling these challenges, aligning with our quantitative results.

Notably, we observe that BYOL and SimSiam produce irrelevant outputs when confronted with changes in viewpoint and occlusion. This tendency may explain their lower recall performance, suggesting a deficiency in learning invariance against occlusion and viewpoint alteration. This insight is crucial as it highlights potential areas for refinement in these models, specifically in enhancing their robustness to such environmental and perspective shifts.



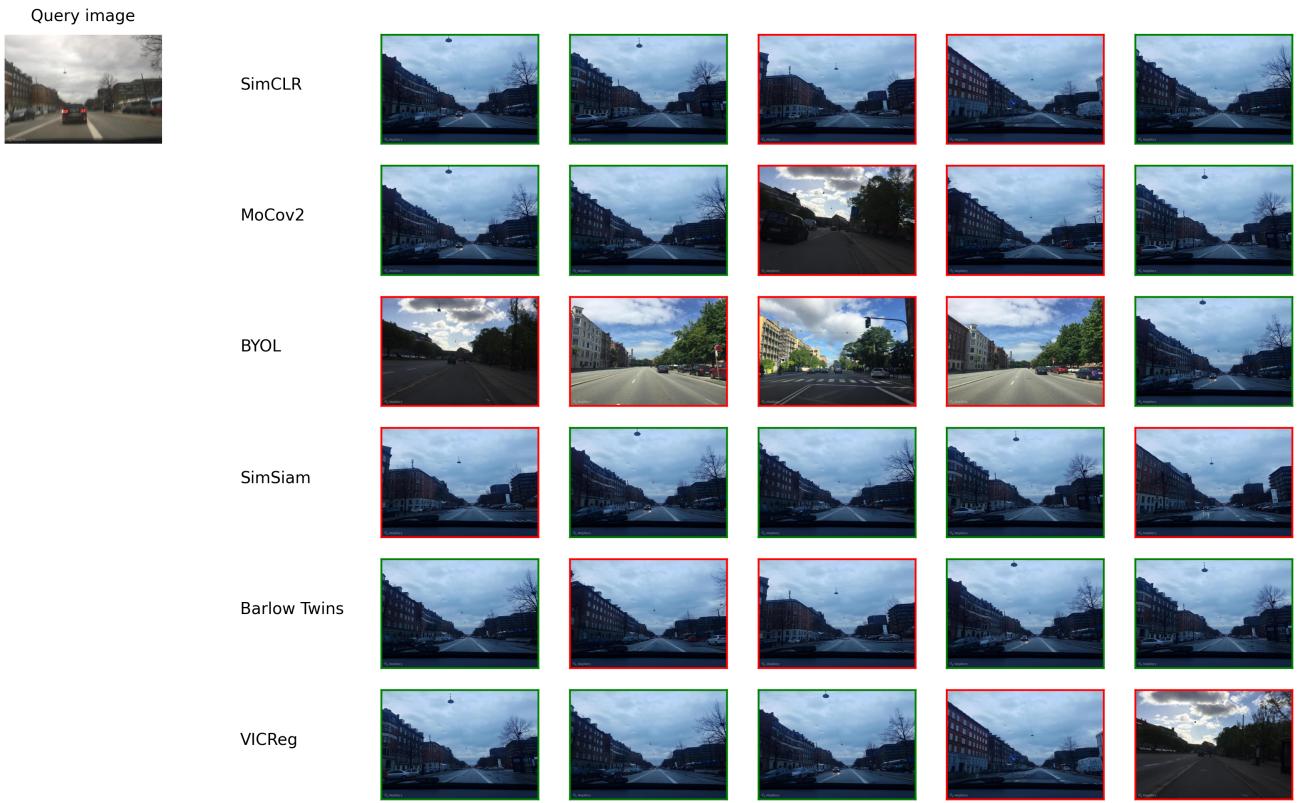


Figure 4. Visualization of top-5 retrieved candidates for **illumination change** across different SSL training strategies

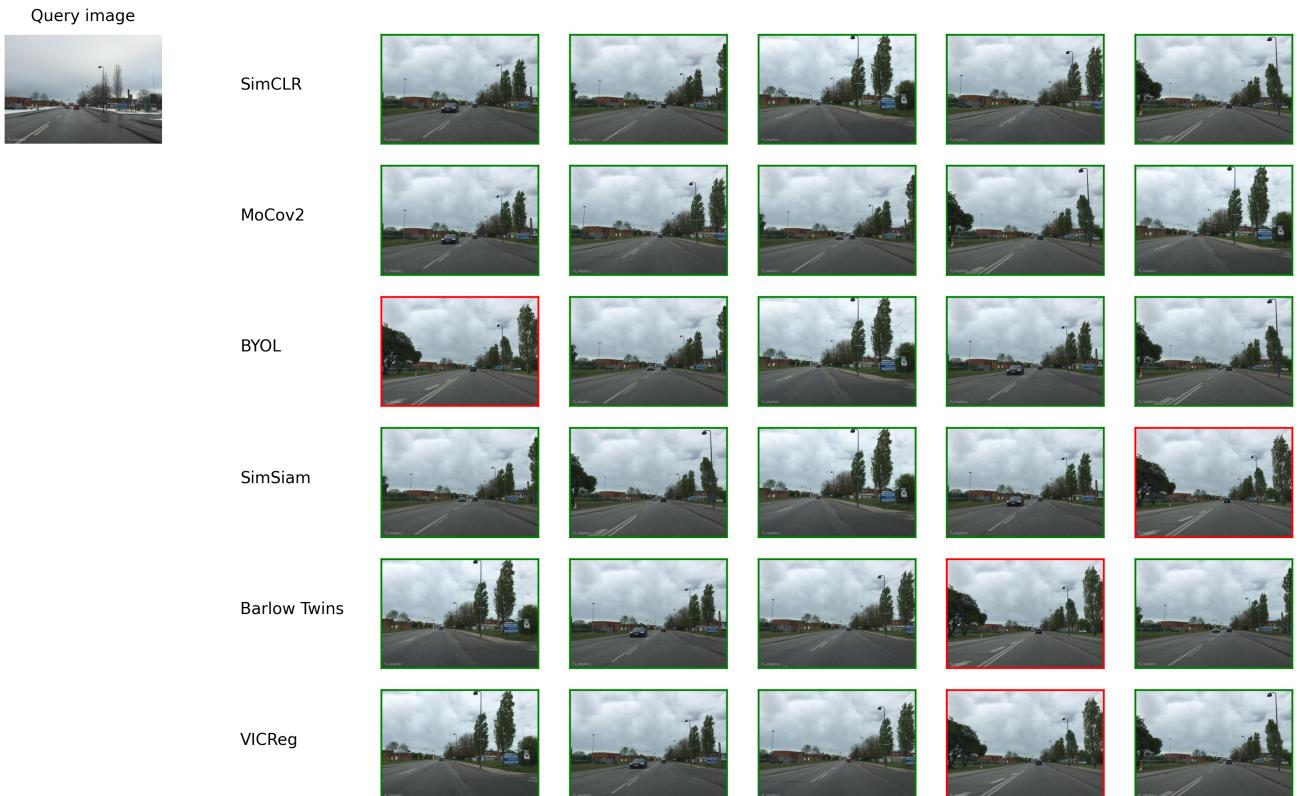


Figure 5. Visualization of top-5 retrieved candidates for **season change** across different SSL training strategies

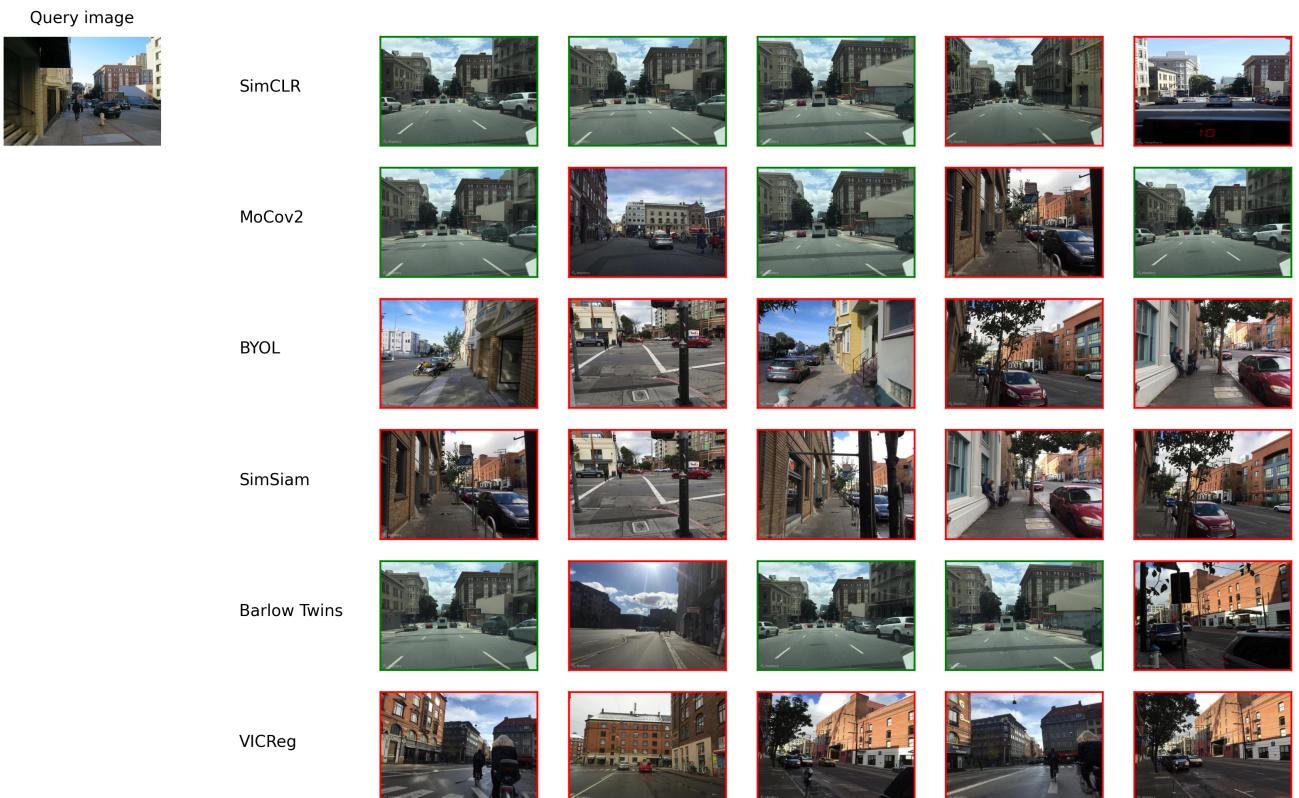


Figure 6. Visualization of top-5 retrieved candidates for **viewpoint change** across different SSL training strategies

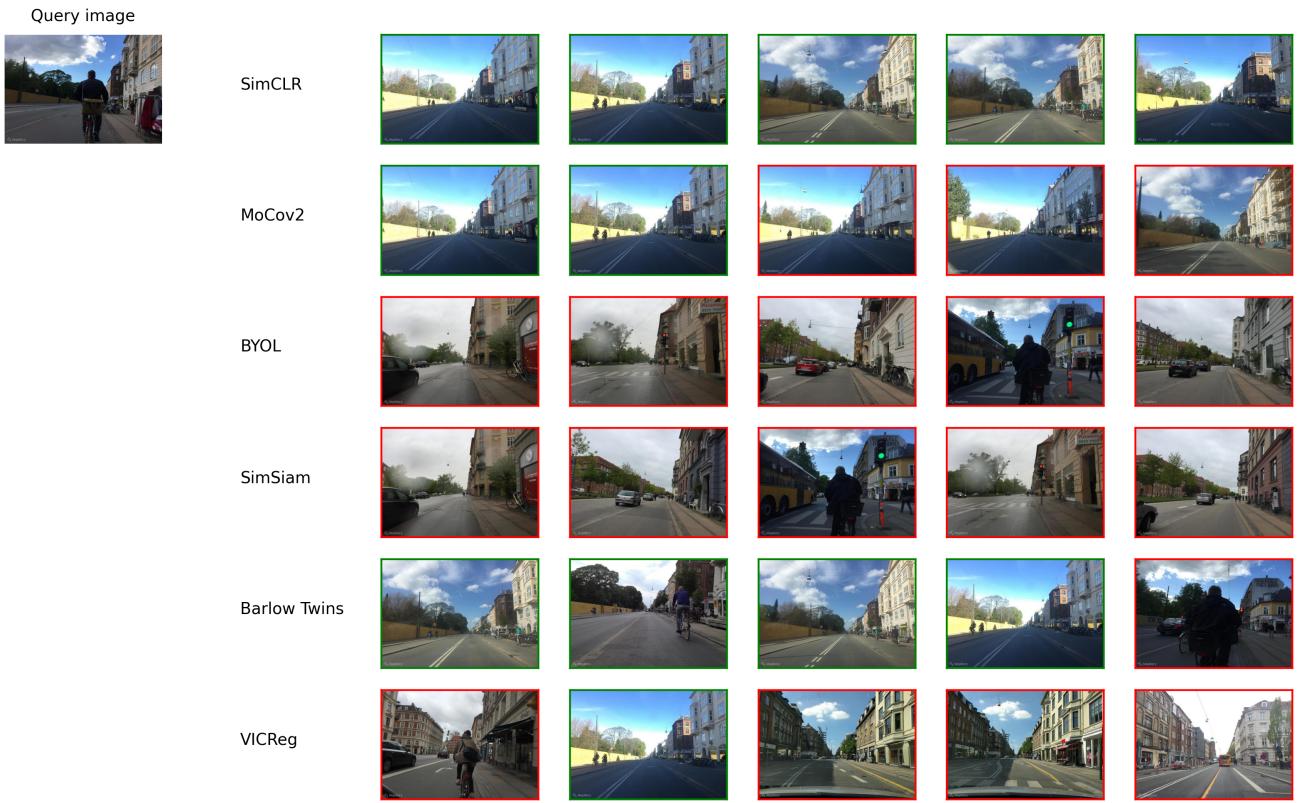


Figure 7. Visualization of top-5 retrieved candidates for **occlusion** across different SSL training strategies