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 optimal 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.

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

	R-SF		
	R@1	R@5	R@10
Our One-Stage Methods with ResNet50-GeM			
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
Our One-Stage Methods with DeiT-S			
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

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.

Table 6. Comparison of state-of-the-art VG methods with our results on large-scale VG datasets. Our models were trained in the MSLS dataset. For the performance in the urban environment (Pitts30k and Tokyo24/7), we further finetuned our models in the Pitts30k dataset. The best results of one-stage and two-stage methods are with **bold** text separately, and the second and third best are underlined. * shows the performance of the first stage without reranking. † shows only the dimensionality of global embeddings but excludes local embeddings.

	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
Two-Stage Methods																
SP-SuperGlue [38]	-	78.1	81.9	84.3	50.6	56.9	58.3	87.2	94.8	**96.4**	88.2	90.2	90.2	29.1	33.5	34.3
Patch-NetVLAD [25]	4096†	79.5	86.2	87.7	48.1	57.6	60.5	88.7	94.5	95.9	86.0	88.6	90.5	44.9	50.2	52.2
TransVPR [47]	256†	86.8	91.2	92.4	63.9	74.0	77.5	89.0	94.9	96.2	79.0	82.2	85.1	58.8	**75.0**	**78.7**
R2Former [56]	256†	89.7	95.0	**96.2**	73.0	85.9	**88.8**	**91.1**	95.2	96.3	**88.6**	**91.4**	**91.7**	**60.6**	66.8	68.7
Our Two-Stage Methods with R2Former																
SimCLR	256†	**90.3**	95.0	95.4	73.2	85.6	88.1	90.5	**95.3**	**96.4**	81.9	87.3	89.5	48.2	54.3	56.3
MoCov2	256†	87.4	93.1	94.1	69.3	84.2	87.0	89.1	94.9	96.1	74.6	82.2	84.8	38.7	47.2	50.0
BYOL	256†	87.2	92.4	93.4	69.1	82.7	84.7	88.3	94.4	95.7	78.1	83.8	86.3	31.1	35.7	37.3
SimSiam	256†	86.2	94.2	95.1	69.3	82.5	85.4	87.9	94.1	95.7	78.7	83.5	84.8	45.0	50.6	52.5
Barlow Twins	256†	89.1	**95.1**	95.9	**73.5**	**86.9**	**88.9**	89.3	94.6	96.3	81.9	87.0	89.8	57.1	65.6	68.3
VICReg	256†	89.7	94.3	96.1	72.6	85.1	88.1	89.4	94.7	96.2	77.5	85.1	87.0	46.8	55.1	57.9
One-Stage Methods																
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	**89.4**	**94.7**	95.9	**85.4**	**91.1**	**93.3**	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	-	-	-
Our One-Stage Methods with ResNet50-GeM																
SimCLR	1024	**84.2**	**92.2**	**94.2**	**63.1**	**78.9**	**83.6**	82.8	91.9	94.6	54.6	74.9	81.9	**39.9**	**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	72.7	85.5	87.7	50.4	66.4	71.4	80.2	91.5	94.4	44.8	63.8	70.8	10.6	18.5	23.5
SimSiam	1024	75.0	85.8	88.6	52.1	67.0	72.2	78.6	89.8	92.7	51.1	67.6	71.4	12.5	21.5	27.0
Barlow Twins	1024	79.5	89.5	91.9	59.2	74.2	79.1	80.8	91.7	94.2	45.7	61.9	70.8	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
Our One-Stage Methods with DeiT-S																
SimCLR	256	81.1	91.1	93.1	58.9	77.1	82.6	84.7	93.9	**96.0**	59.4	76.2	80.0	24.9	38.9	46.1
MoCov2	256	76.1	88.5	91.1	56.8	75.2	78.7	80.8	92.4	95.0	50.8	69.8	77.1	15.4	26.4	33.0
BYOL	256	58.2	75.3	79.6	37.7	54.0	60.4	76.6	89.4	92.9	43.2	62.2	68.6	4.1	7.9	10.6
SimSiam	256	56.2	76.2	80.1	35.3	52.3	58.7	79.7	91.0	93.6	47.3	63.8	74.0	6.2	11.5	15.4
Barlow Twins	256	79.7	91.4	93.1	59.1	76.1	81.5	82.6	92.1	95.0	58.4	75.2	80.6	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

Table 7. Comparison of validation performance in MSLS dataset for the different training stages of our two-stage methods.

R2Former [56]			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	R@1	R@5	R@10
Global Retrieval Training																				
79.3	90.8	92.6	81.1	91.1	93.1	76.1	88.5	91.1	58.2	75.3	79.6	56.2	76.2	80.1	79.7	91.4	93.1	75.8	89.5	91.9
Reranking Training																				
88.4	93.4	94.9	89.2	94.3	95.4	86.8	93.1	93.9	81.2	86.2	87.4	83.0	87.3	88.8	88.2	92.7	93.6	88.5	92.7	93.8
End-to-end Finetuning																				
89.7	95.0	96.2	90.3	95.0	95.4	87.4	93.1	94.1	87.2	92.4	93.4	86.2	94.2	95.1	89.1	95.1	95.9	89.7	94.3	96.1

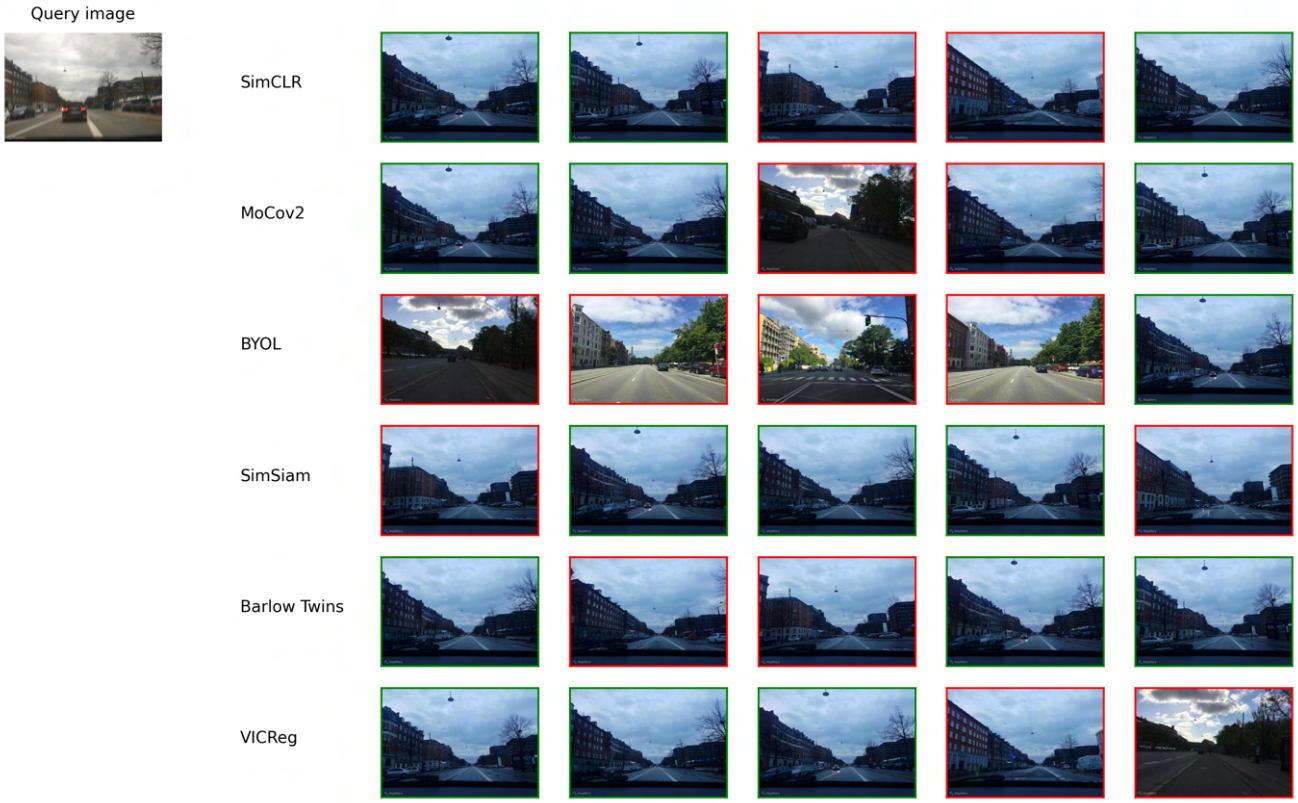

Figure 4. Visualization of top-5 retrieved candidates for **illumination change** across different SSL training strategies

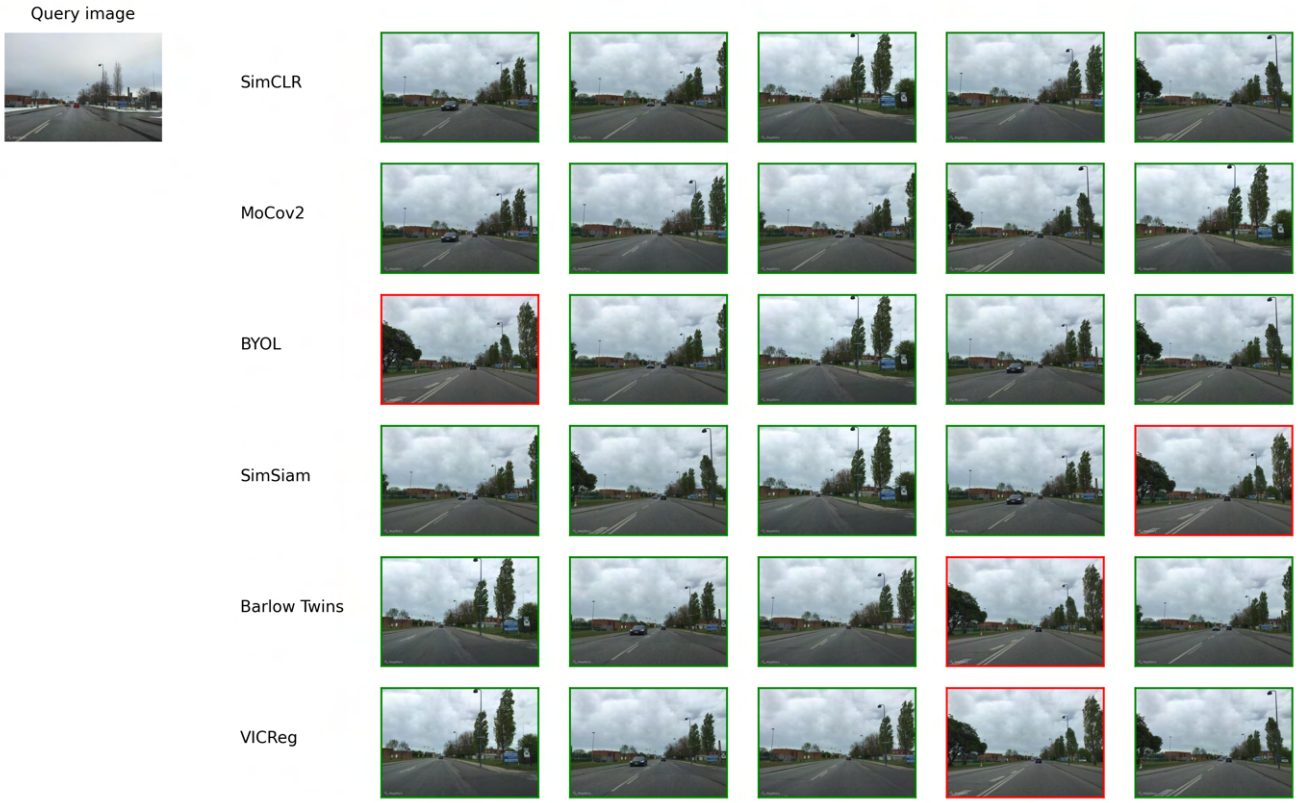

Figure 5. Visualization of top-5 retrieved candidates for **season change** across different SSL training strategies

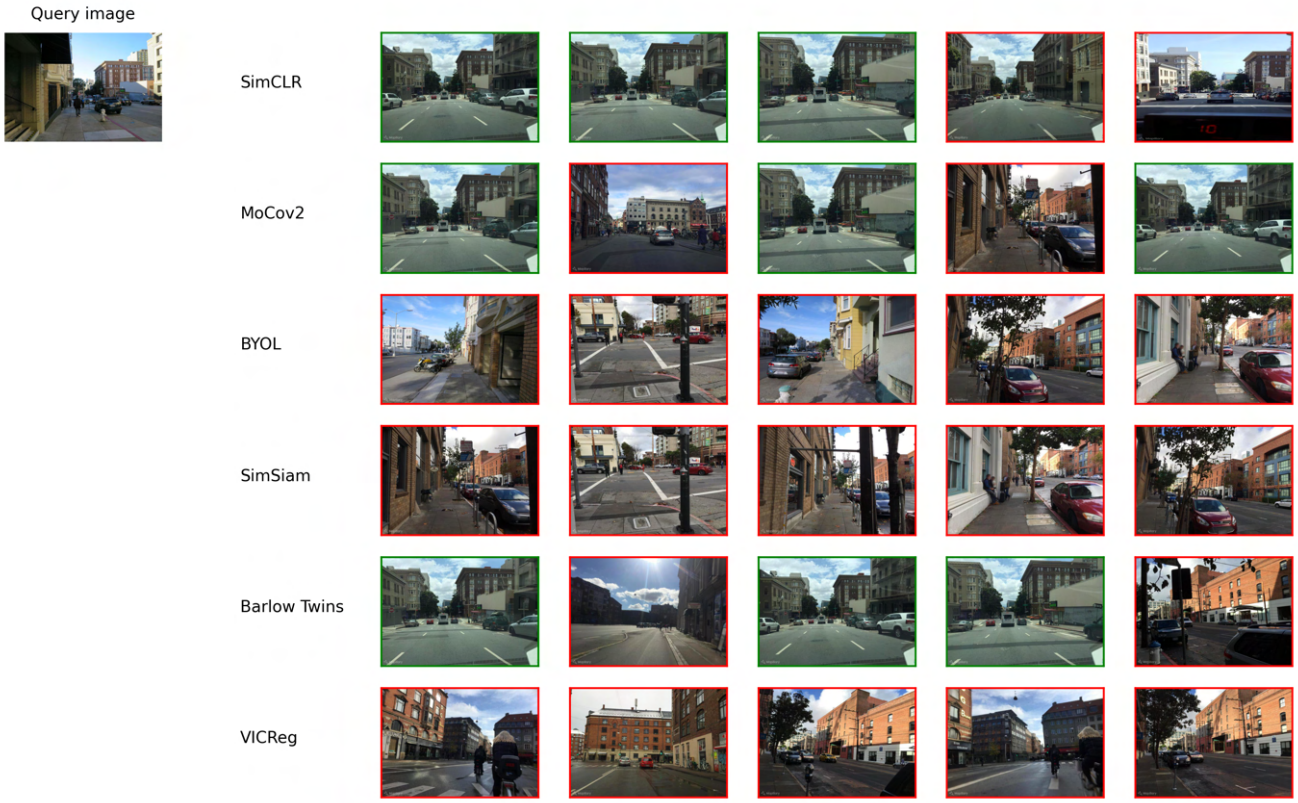

Figure 6. Visualization of top-5 retrieved candidates for **viewpoint change** across different SSL training strategies

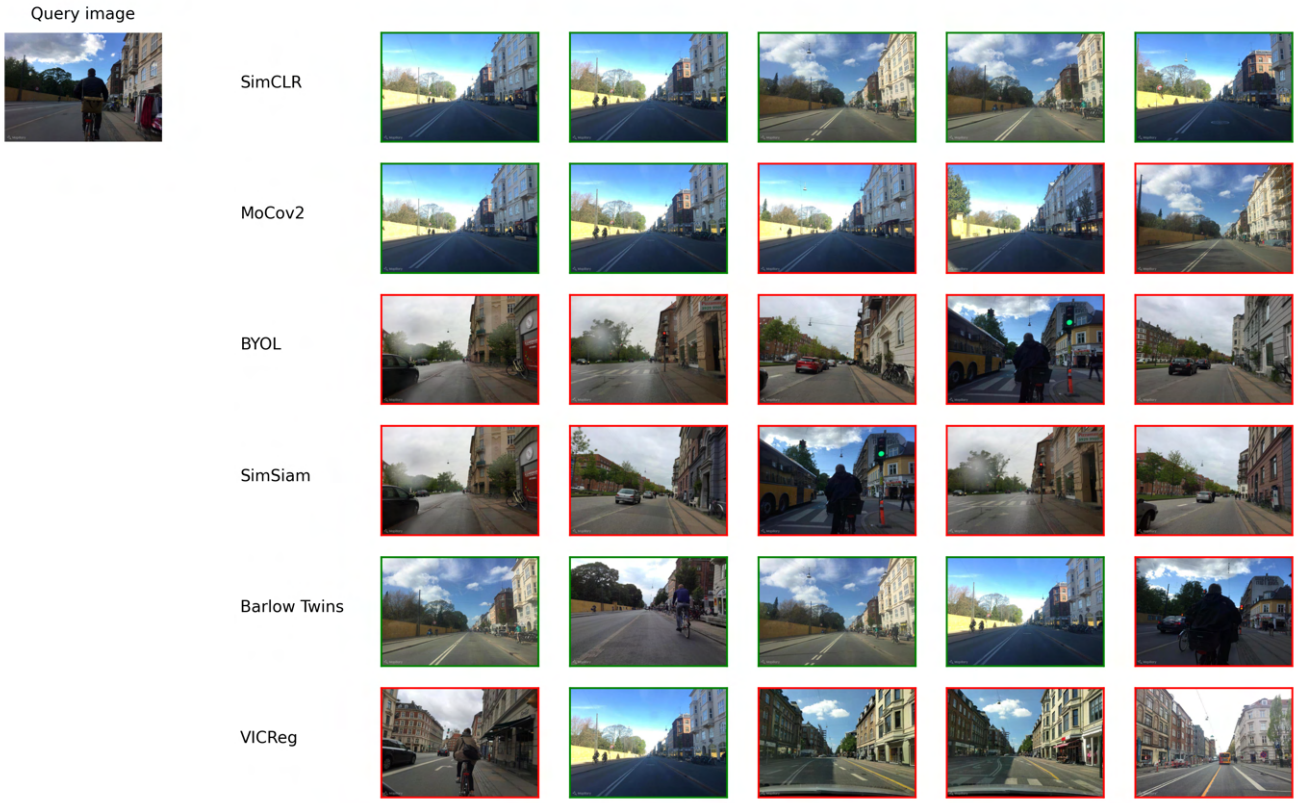

Figure 7. Visualization of top-5 retrieved candidates for **occlusion** across different SSL training strategies