In-context adaptation of place recognition through self-supervised learning from video

Kiavash Jamshidi Gülhan Şikaroğlu Hermann Blum

***Abstract*— This paper studies in-domain adaptation of the video-pretrained DoRA backbone for Visual Place Recognition (VPR). In SLAM we only need to tell apart places the robot has actually seen, not every possible image, raising the question: can domain-specific descriptors beat strong generic ones, and how do we learn them without labels? Starting from the released generic-video checkpoint, we adapt DoRA to each traversal's videos using a jointly trained SALAD aggregator and a weighted multi-loss objective. Across four benchmark datasets (indoor, urban, suburban, day/night, continuous video), in-domain DoRA+SALAD improves Recall@5/10, with the largest gain on St Lucia (R@5: +16%, R@10: +22%), showing that contextualizing a video-pretrained backbone on the deployment domain is key. Our weighted multi-objective loss jointly leverages temporal consistency (DoRA distillation) and appearance-based retrieval (SALAD InfoNCE), driving the end-to-end in-domain optimization. Frozen optimal-transport aggregation on unadapted DoRA features reduces recall, underscoring the need for this joint training.**

I. Introduction

Visual Place Recognition (VPR) is a long-studied problem in robotics and computer vision [1], [2]: given a query image, identify the location by retrieving the most similar image from a geo-referenced database. VPR is a core module in autonomous navigation and localization pipelines, including SLAM systems [3], where robust place recall is required despite strong changes in viewpoint, illumination, and seasonal appearance.

Robots repeatedly see the same places along their routes; a natural strategy is to learn place descriptors *from those past observations* so they are tuned to the deployment environment. We therefore ask: do domain-adapted descriptors beat strong generic ones? We start from generic self-supervised backbones (DINO/DINOv2 image pretraining, DoRA video pretraining) and then adapt DoRA to each traversal's video using self-supervision plus a retrieval loss.

Our work is motivated by the observation that VPR failures in practice are often rather caused by domain shift (lighting, season, camera pose) rather than model capacity. Adapting on the robot's own past observations lets us align the backbone to the exact viewpoint/appearance statistics it will face, while the approach is entirely self-supervised. To this end, we adopt the DoRA [4] self-supervision objective to learn a domain-adapted feature backbone. We then experiment with different feature pooling strategies and how to adapt these in a self-supervised way.

We evaluate our approach on several target datasets (St Lucia, Oxford, Gardens) and find that adaptation always improves in these environments compared to the DoRA baseline. In the challenging St Lucia, where DINO/DINOv2 struggle to distinguish images, our adapted descriptor achieves the best results overall. Figure 1 shows a night-to-day query where off-the-shelf DoRA+GeM fails, but in-domain DoRA+SALAD succeeds.

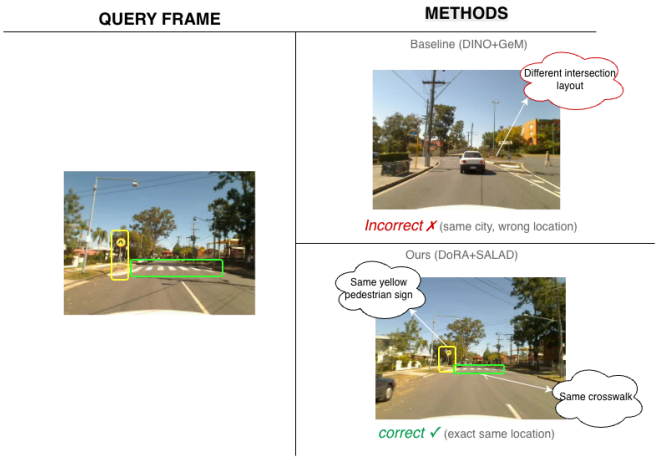

Fig. 1: Qualitative example: query (left), top-1 from DINO (middle, incorrect), top-1 from in-context DoRA+SALAD (right, correct).

In summary, our contributions are:

- A self-supervised in-context learning pipeline for VPR: domain-adapt the video-pretrained DoRA backbone on target traversal videos without labels, and optionally co-train SALAD aggregation end-to-end.
- Extensive evaluation and ablations across diverse datasets (indoor, urban, suburban, day/night, continuous video), comparing generic image SSL (DINO/DINOv2), generic video SSL (DoRA), in-context fine-tuning, and aggregation variants (GeM vs. SALAD/CriSALAD, frozen vs. joint).

II. Related Work

Self-supervised representation learning. Contrastive learning methods such as SimCLR [5] and MoCo [6] established that strong visual representations can be learned from unlabelled data by maximizing agreement between augmented views of the same image, with objectives related to supervised contrastive loss [7]. DINO [8] adapts this paradigm to Vision Transformers [9] through a teacher-student self-distillation scheme, producing emergent semantic attention maps without explicit supervision. DINOv2 [10] scales this approach to a larger curated dataset, delivering features that generalize across segmentation benchmarks

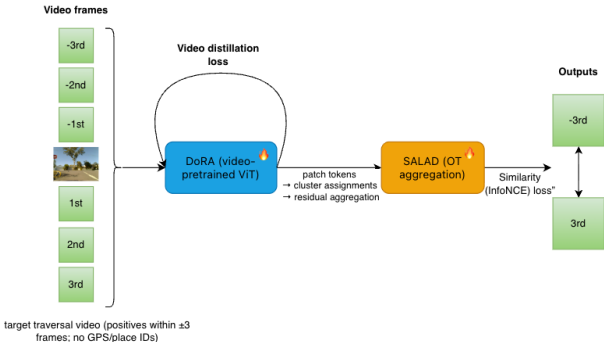

Fig. 2: In-context training pipeline. Target traversal frames (query, temporal positive, negatives) pass through the DoRA video-pretrained ViT; a video distillation loss keeps the backbone stable. Patch tokens are aggregated by SALAD via Sinkhorn OT into a global descriptor optimized with an InfoNCE similarity loss.

Fig. 3: Gardens Point: both DINO (middle) and in-context DoRA+SALAD (right) retrieve the correct place for the query (left), showing parity on this dataset.

such as ADE20K [11], object detection, and image retrieval. DoRA [4] departs from image collections entirely: it trains on a single long unlabelled video by tracking attention-based object regions across frames and using their temporal consistency as a self-supervised signal. While DoRA was evaluated on semantic segmentation [11] and showed competitive performance, its behavior in retrieval-oriented tasks such as VPR remains unexplored.

Visual Place Recognition and descriptor aggregation. Early VPR relied on appearance-based methods such as FAB-MAP [12] and sequential matching approaches [13]. Learning-based methods transformed the field: NetVLAD [14] introduced trainable cluster-based pooling, while approaches combining deep local and global features [15] and multi-scale descriptors such as PatchNetVLAD [16] further improved robustness. Recent work focuses on scalable training pipelines [17], viewpoint-robust models [18], and feature mixing strategies [19]. AnyLoc [20] shows that DINOv2 features generalize across diverse VPR environments without dataset-specific retraining. SALAD [21] replaces soft VLAD assignment with a Sinkhorn-normalized optimal transport problem and adds a learnable dustbin cluster for uninformative patches, achieving state-of-the-art recall. CriSALAD [22] extends SALAD with cross-image descriptor relationships during training. Both methods rely on image-pretrained backbones and do not explore video-pretrained encoders.

Domain shift in VPR. Night-time and seasonal appearance changes remain major challenges. Methods such as night-to-day image translation [23] address this by creating paired training data for domain adaptation. These studies confirm that many VPR failures stem from domain mismatch, motivating the evaluation of robust backbones under varied conditions.

III. Method

Figure 2 summarizes our in-domain adaptation pipeline: a generic-video DoRA backbone is fine-tuned on the target traversal videos while training SALAD with a weighted multi-loss.

We start by training on the released generic-video DoRA checkpoint. For each target dataset we extract its traversal video frames (40% test split held out) and then optimize DoRA and SALAD end-to-end. GeM on unadapted DoRA serves as the generic baseline because GeM is a strong, well-understood pooling head for global retrieval. SALAD-on-frozen-DoRA is the non-adapted aggregation baseline because it isolates the benefit of optimal-transport aggregation without changing the backbone.

A. Training Objective

When SALAD is applied to frozen DoRA features, the backbone cannot adapt to the clustering objective. We resolve this mismatch by jointly training DoRA and SALAD using a weighted multi-loss objective:

$$\mathcal{L} = \lambda_{\text{ret}}\,\mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{dist}}\,\mathcal{L}_{\text{distill}}, \tag{1}$$

where $\mathcal{L}_{\text{InfoNCE}}$ is the SALAD retrieval loss on mined positive and hard-negative pairs, and $\mathcal{L}_{\text{distill}}$ preserves the original DoRA self-supervised signal. Gradients of $\mathcal{L}_{\text{ret}}$ are backpropagated into the DoRA backbone, allowing patch representations to adapt specifically for clustering-based aggregation.

a) Self-distillation details.: We keep the standard DoRA/DINO self-distillation: an EMA teacher (momentum 0.996→1.0) supervises the student on video crops (1 global + 8 local). The teacher sees only the global crops; student augments include jitter, grayscale, blur, flip, and local 112px crops as listed in Training details. Distillation loss is unchanged from DoRA; we reuse DINO's hyperparameters because they are stable for ViT-S and keep the focus on the effects of in-domain adaptation. We simply run it alongside the SALAD InfoNCE loss (weights λ_{ret}=1.0, λ_{dist}=1.0) in the joint pipeline.

TABLE I: Main results contrasting off-the-shelf pretrained features with in-context (target video) adaptation. DoRA (none) uses the released generic-video checkpoint with GeM. DoRA in-context uses the same checkpoint adapted on target videos with GeM. DoRA aggregated (ours) uses joint DoRA+SALAD fine-tuning on target videos.

		UQ St Lucia			Oxford RobotCar			QUT Gardens			17-Places		
features	adaptation	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
DINO	none	0.233	0.367	0.433	0.673	1.000	1.000	0.870	0.970	1.000	0.155	0.310	0.539
DINOv2	none	0.200	0.400	0.533	0.469	1.000	1.000	0.815	0.980	0.995	0.552	0.721	0.770
DoRA	none	0.033	0.200	0.300	0.204	0.653	0.857	0.345	0.595	0.720	0.526	0.621	0.655
DoRA	in-context video	0.287	0.536	0.613	0.450	0.841	0.935	0.405	0.615	0.715	–	–	–
DoRA + SALAD (frozen DoRA)	in-context video (SALAD only)	0.169	0.414	0.530	0.399	0.777	0.862	0.265	0.505	0.675	–	–	–
DoRA aggregated (ours)	in-context video	**0.301**	**0.620**	**0.751**	**0.477**	**0.864**	**0.951**	**0.350**	**0.650**	**0.735**	–	–	–

IV. Experiments

A. Methods Compared

We cover three questions: (i) how strong are generic image SSL baselines (DINO, DINOv2) when paired with a simple GeM head; (ii) what do we gain from video pretraining alone (off-the-shelf DoRA+GeM) versus in-context backbone tuning (DoRA+GeM adapted on target videos); and (iii) how much does optimal-transport aggregation help, both when the backbone is frozen and when it is co-trained. Accordingly, we compare six settings: 1) DINO + GeM (generic image SSL), 2) DINOv2 + GeM (stronger generic image SSL), 3) DoRA + GeM off-the-shelf (generic video SSL, no target training), 4) DoRA + GeM in-context (backbone fine-tuned on target videos), 5) DoRA + SALAD (frozen DoRA, SALAD trained on target videos), 6) DoRA + SALAD joint (ours: backbone and SALAD co-trained on target videos).

B. Datasets and In-context Video Matching

We evaluate on four VPR benchmarks: **UQ St Lucia**, **Oxford RobotCar**, **QUT Gardens Point**, and **17-Places**. For in-context adaptation, we train on the training/validation portion (40% test held out) of each dataset's traversal videos; positives are defined within ± 1 and ± 3 temporal windows, matching the evaluation protocol, no place IDs or GPS labels are used. Splits are carved as contiguous temporal blocks (e.g., first 60% train, next 20% val, last 20% test) so no adjacent frames cross split boundaries, preventing near-duplicate leakage. The generic DoRA checkpoint is fine-tuned on these target frames.

a) Training details: ViT-S/16 with EMA teacher $0.996 \rightarrow 1.0$, FP16, AdamW with cosine weight decay. We train 20 epochs with 10-epoch warmup, cosine LR to 10^{-6} (peak from $5 \times 10^{-4} \cdot \mathrm{bs}/256$). Losses: $\lambda_{\text{ret}}=1.0$, $\lambda_{\text{dist}}=1.0$, SALAD temp 0.07. Batches: 16 (DINO), 64 (SALAD). Positives: ± 3 temporal window; negatives: in-batch only. Augmentations: multi-scale global/local crops, flip, color jitter, grayscale, blur; SALAD branch uses resize 256, center-crop 224, and ImageNet normalization. SALAD uses 256 clusters. Splits keep a 40% held-out test set with no temporal adjacency across splits.

Note. DoRA fine-tunes on 8-frame clips sampled every 60 frames (no $\pm k$ labels). The ± 3 temporal window applies to SALAD/joint pairing and to evaluation (a retrieval is correct within ± 3). Wider windows or cross-traversal positives remain future work.

TABLE II: SALAD/CriSALAD variants with frozen DoRA and joint DoRA+SALAD on 40% test splits.

Dataset	Training	Feature Aggregation	R@1	R@5	R@10
UQ St Lucia	1st DoRA, 2nd aggregation	GeM (baseline)	0.287	0.536	0.613
	1st DoRA, 2nd aggregation	SALAD (frozen)	0.169	0.414	0.530
	1st DoRA, 2nd aggregation	SALAD + low-light aug.	0.188	0.434	0.544
	1st DoRA, 2nd aggregation	CriSALAD (frozen)	0.188	0.417	0.528
	1st DoRA, 2nd aggregation	CriSALAD + aug.	0.138	0.337	0.445
	1st DoRA, 2nd aggregation	SALAD (± 3 window)	0.199	0.470	0.591
	joint DoRA + aggregation	SALAD (joint)	**0.301**	**0.620**	**0.751**
QUT Gardens	1st DoRA, 2nd aggregation	GeM (baseline)	0.405	0.615	0.715
	1st DoRA, 2nd aggregation	SALAD (frozen)	0.265	0.505	0.675
	1st DoRA, 2nd aggregation	SALAD + low-light aug.	0.280	0.560	0.685
	1st DoRA, 2nd aggregation	CriSALAD (frozen)	0.290	0.530	0.660
	1st DoRA, 2nd aggregation	CriSALAD + aug.	0.240	0.525	0.665
	1st DoRA, 2nd aggregation	SALAD (± 3 window)	0.340	0.590	0.710
	joint DoRA + aggregation	SALAD (joint)	0.350	**0.650**	**0.735**
Oxford RobotCar	1st DoRA, 2nd aggregation	GeM (baseline)	0.450	0.841	0.935
	1st DoRA, 2nd aggregation	SALAD (frozen)	0.399	0.777	0.862
	1st DoRA, 2nd aggregation	SALAD + low-light aug.	0.394	0.775	0.861
	1st DoRA, 2nd aggregation	CriSALAD (frozen)	0.384	0.764	0.850
	1st DoRA, 2nd aggregation	CriSALAD + aug.	0.261	0.593	0.692
	1st DoRA, 2nd aggregation	SALAD (± 3 window)	0.423	0.798	0.882
	joint DoRA + aggregation	SALAD (joint)	**0.477**	**0.864**	**0.951**
17-Places	1st DoRA, 2nd aggregation	GeM (baseline)	0.517	0.625	0.579
	1st DoRA, 2nd aggregation	SALAD (frozen)	0.561	0.681	0.800
	1st DoRA, 2nd aggregation	SALAD + low-light aug.	0.488	0.655	0.782
	1st DoRA, 2nd aggregation	CriSALAD (frozen)	0.495	0.626	0.751
	1st DoRA, 2nd aggregation	CriSALAD + aug.	0.536	0.683	0.771
	1st DoRA, 2nd aggregation	SALAD (± 3 window)	0.560	0.701	0.805
	joint DoRA + aggregation	SALAD (joint)	0.467	**0.736**	**0.853**

C. Main Results: Pretrained vs. In-context DoRA

Table I highlights the workshop focus: generic pretrained descriptors vs. in-context DoRA fine-tuning, and the additional gain from joint DoRA+SALAD. DoRA in-context (GeM) improves over off-the-shelf DoRA; joint DoRA+SALAD is best on all three adapted datasets.

a) Main results discussion.: DINO/DINOv2 remain very strong generic baselines: DINO leads on Gardens and Oxford, DINOv2 on heterogeneous scenes. Generic DoRA trails DINO on all datasets, especially St Lucia. Adapting DoRA in-context (GeM) narrows the gap, and joint DoRA+SALAD surpasses DINO on St Lucia and clearly improves over the non-adapted DoRA baseline on Gardens/Oxford, showing the gain comes from adapting the video-pretrained backbone to target videos. We do *not* adapt DINO, so the improvement is attributable to in-context video fine-tuning of DoRA (and SALAD), not to extra supervision.

Off-the-shelf DoRA trails DINO/DINOv2 on St Lucia and RobotCar, so we test whether SALAD-based aggregation can recover that gap.

D. Ablation: SALAD Variants and Joint Training

Frozen SALAD/CriSALAD underperform GeM, while joint DoRA+SALAD consistently restores and surpasses baseline recall; therefore we select joint training as the final

method. CriSALAD in theory benefits from cross-image assignments, but in our small, single-traversal batches (±3 window) the extra coupling likely overfits or is underutilized; its entropy/dustbin settings also were not re-tuned for this low-diversity regime.

E. Discussion

Three observations emerge. First, video-based self-supervision does not automatically produce better VPR representations than large-scale image pretraining: temporal consistency encourages viewpoint invariance, but this alone does not guarantee discriminative global descriptors across diverse scenes. Second, optimal-transport aggregators are sensitive to the feature distribution of the backbone; when this distribution does not support balanced cluster assignment, Sinkhorn normalization hurts rather than helps. Third, joint optimization is an effective remedy: by co-adapting feature extraction and aggregation, the backbone produces representations that satisfy the structural requirements of clustering-based descriptor learning.

a) Limitations and practical notes.: Our in-context training uses only the available traversal videos; no GPS or place IDs are leveraged. This keeps annotation cost zero but limits how precisely positives can be mined, temporal ± 3 windows may include hard-but-correct views as negatives. Future work will evaluate larger temporal windows, multi-traversal batching, and mixed-modality positives (e.g., LiDAR frames) to tighten positive mining without supervision.

V. Conclusion

We focused this workshop submission on in-domain adaptation of the video-pretrained DoRA backbone: starting from the generic-video checkpoint, we fine-tuned DoRA together with SALAD on the target traversal videos. Across four VPR datasets, frozen optimal-transport aggregation on unadapted DoRA reduced recall, while in-domain joint training consistently improved Recall@5/10 (and Recall@1 on two datasets). This highlights that the key lever is contextualizing a video-pretrained backbone to deployment data, aligning with the workshop emphasis on rigorous, in-context perception. These findings show that aggregation method selection cannot be decoupled from backbone training strategy. Future work will investigate whether these joint-optimization benefits extend to other video-pretrained encoders and to multi-modal VPR settings incorporating additional sensor modalities.

References

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *ieee transactions on robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[2] M. Zaffar, S. Garg, M. Milford, J. Kooij, D. Flynn, K. McDonald-Maier, and S. Ehsan, "Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2136–2174, 2021.

[3] N. Sünderhauf, *Switchable Constraints for Robust Simultaneous Localization and Mapping and Satellite-Based Localization*. Springer Nature, 2023, vol. 137.

[4] S. Venkataramanan, M. N. Rizve, J. Carreira, Y. M. Asano, and Y. Avrithis, "Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video," *arXiv preprint arXiv:2310.08584*, 2023.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[7] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[10] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[11] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International journal of computer vision*, vol. 127, no. 3, pp. 302–321, 2019.

[12] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *The International journal of robotics research*, vol. 27, no. 6, pp. 647–665, 2008.

[13] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *2012 IEEE international conference on robotics and automation*. IEEE, 2012, pp. 1643–1649.

[14] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[15] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *European conference on computer vision*. Springer, 2020, pp. 726–743.

[16] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 141–14 152.

[17] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4878–4888.

[18] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 080–11 090.

[19] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "Mixvpr: Feature mixing for visual place recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 2998–3007.

[20] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1286–1293, 2023.

[21] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2024, pp. 17 658–17 668.

[22] J. Xu, Y. Ming, M. Xu, Y. Fan, Y. Zhang, and W. Kong, "Crisalad: Robust visual place recognition using cross-image information and optimal transport aggregation," *Applied Sciences*, vol. 15, no. 10, p. 5287, 2025.

[23] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool, "Night-to-day image translation for retrieval-based localization," in *2019 International conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 5958–5964.