Distillation Improves Visual Place Recognition for Low Quality Images

Anbang Yang[1*], Ge Jin[1*], Junjie Huang[1], Yao Wang[1], John-Ross Rizzo[2], Chen Feng[1✉]
New York University[1] NYU Grossman School of Medicine[2]
{ay1620, gj2148, jh7956, yaowang, cfeng}@nyu.edu[1], JohnRoss.Rizzo@nyulangone.org[2]

Abstract

Real-world applications of Visual Place Recognition (VPR) often rely on cloud computing, where query images or videos are transmitted elsewhere for visual localization. However, limited network bandwidth forces a reduction in image quality, which degrades global image descriptors and consequently VPR accuracy. We address this issue at the descriptor extraction level with a knowledge-distillation framework that transfers feature representations from high-quality images to low-quality ones, allowing VPR methods to produce more discriminative descriptors from the latter. Our approach leverages three complementary loss functions —- the Inter-channel Correlation Knowledge Distillation (ICKD) loss, Mean Squared Error (MSE) loss, and a weakly supervised Triplet loss —– to guide a student network in approximating the high-quality features produced by a teacher network. Extensive experiments on multiple VPR methods and datasets subjected to JPEG compression, resolution reduction, and video quantization demonstrate significant improvements in recall rates. Furthermore, our work fills a gap in the literature by exploring the impact of video-based data on VPR performance, thereby contributing to more reliable place recognition in resource-constrained environments. Source code and data are available at https://ai4ce.github.io/LoQI-VPR/

1. Introduction

Visual Place Recognition (VPR) has rapidly become a foundational component in machine perception. Through learning compact vector representations of RGB images, or global image descriptors, VPR methods localize unseen images by recalling one or more known locations among a database of images based on descriptor similarity [11, 24]. VPR is critical for real-world applications ranging from robotic navigation in complex industrial environments to augmented reality systems and way finding aids for vulnerable populations [8, 39, 41].

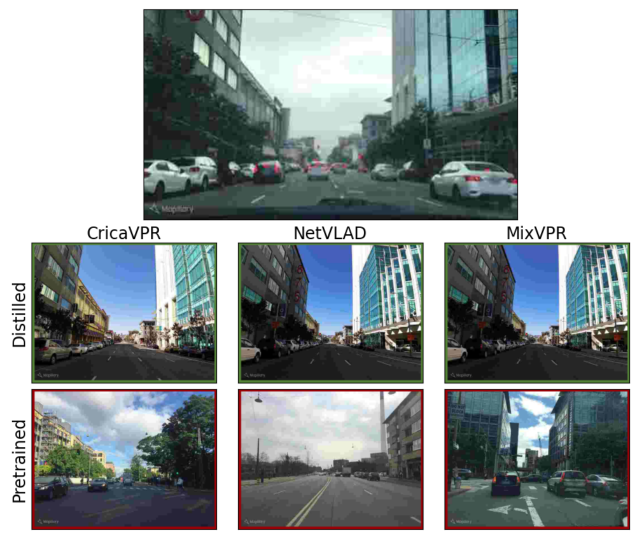

Figure 1. **Enhanced Low Quality Image Retrieval:** After knowledge distillation from high quality images, this JPEG-compressed query image from the Mapillary SLS dataset (top center) is correctly localized by multiple VPR methods (top row), whereas they failed beforehand (bottom row).

However, VPR deployment at scale comes with an inherent dilemma between VPR performance and computational / networking costs. On one hand, state-of-the-art (SOTA) VPR methods leverage deep learning to produce highly discriminative and generalizable global descriptors [2, 23]. However, such methods are computationally costly for both descriptor extraction and matching [10], making large-scale deployment on consumer / edge devices impractical. Besides computational cost, there is also a further

[*] Equal contribution.
✉ Corresponding author (cfeng@nyu.edu). This work is supported partly by NSF Grants 2238968 and 2345139; by the National Eye Institute and Fogarty International Center under Grants R21EY033689, R33EY033689, and R01EY036667; and by the NYU IT High Performance Computing resources, services, and staff expertise.

dilemma caused by the storage demands of the VPR image database, which is difficult to maintain locally given the rapid growth in scene diversity and volume for real-world VPR applications [5, 19].

Given these challenges, one option is to perform VPR locally on low-end devices using methods specifically designed for low computational cost. The second option is to perform VPR over the cloud, where images are transmitted elsewhere for centralized processing. Arguably, the cloud VPR approach has greater potential for accurate localizations, as it carries the aforementioned benefits of SOTA VPR methods. However, network bandwidth constraints are common when many devices simultaneously capture images to query the VPR database, hampering real-time communication with servers. As most real-world VPR applications localize over image sequences, the underlying video streaming service could alleviate bandwidth pressures by reducing resolution and increasing image or video compression. However, such quality reductions to VPR methods' visual inputs significantly hamper localization accuracy [32].

Our work offers a potential solution to the dilemmas of real-world VPR, benefiting from the performance and generalizability of SOTA VPR methods while mitigating VPR accuracy loss from lowered image quality. As networking-induced image quality reduction within the cloud VPR paradigm is not easily eliminated, we propose a knowledge distillation framework to improve the performance of VPR methods under low-quality images. Although image processing techniques not specific to VPR such as super-resolution [26, 42] and image deblurring [18, 20] can potentially restore images' visual contents after network transmission, these methods do not account for images' role as place representations and thus are not directly beneficial for VPR performance. In short, we offer the following contributions:

- Utilizing knowledge distillation, we develop a training framework to improve the localization performance of VPR global descriptors extracted from low-quality images. Our method is not limited to any particular VPR method.
- We validate our framework across multiple public VPR datasets and various types of low-quality images, including JPEG compression, resolution reduction, and video quantization. In doing so, we also curate a video-based VPR dataset to address the low availability of video data for VPR research.

2. Related Work

2.1. VPR Under Low-Quality Image Conditions

Prevailing VPR literature frames the problem of image-based localization as an image retrieval task, and research in VPR methods has traditionally focused on improving retrieval accuracy under high-level visual changes such as illumination changes, seasonal variation, and viewpoint shifts [24]. However, the impact of *image quality degradation* (e.g., low resolution, heavy compression, or motion blur) has received comparatively less attention. Recent studies have begun to highlight this gap. For instance, the accuracy of deep learning-based VPR methods such as NetVLAD [3] suffers significantly from low-resolution images [32]. In addition to lowered image resolution, increased JPEG compression has also been shown to negatively impact VPR methods [31]. Given the high compatibility and widespread adoption of the JPEG image format [14], it is prevalent in image transmission when deploying VPR at scale [40]. Furthermore, many widely used VPR datasets for studying the aforementioned high-level retrieval challenges also store images in JPEG format. Examples include Tokyo247 [35] (illumination changes), Mapillary SLS [38] (large temporal variation), Nordland [30] (seasonal variation), and GSV-Cities [1] (viewpoint shifts). Although in our work we consider multiple forms of image quality reduction, we place a special emphasis on JPEG compression for its relevance to real-world deployments and existing VPR datasets.

Besides image quality reduction, the cloud VPR paradigm also faces the challenge of video compression. Within VPR research, localization accuracy under video compression and specifically video quantization is not extensively studied. Even lightweight VPR methods [4, 9, 10], the alternative solution to the dilemmas of computational cost / storage demands introduced in 1, still focus on the image domain. As video quantization introduces additional artifacts that further degrade image quality, we seek to better understand video compression's effects on VPR descriptor extraction. To our best knowledge, we also contribute towards a literature gap on publicly available datasets offering video data.

2.2. Knowledge Distillation and Transfer Learning in VPR

Likewise, research on effective methods to extract more representative global descriptors from low-quality images is lacking in the VPR community. Existing solutions have addressed the issue at the descriptor matching and image-retrieval level. The authors of [31] propose a sequence-based image-retrieval strategy [33], which queries the database images multiple times with lower per-query cost thanks to JPEG compression. Furthermore, predicting the retrieval success for each query image could favor points within the sequential queries that are more easily localized [7]. We instead wish to explore knowledge distillation as a means to directly improve descriptor extraction and thus increase the accuracy of database retrieval via individual queries.

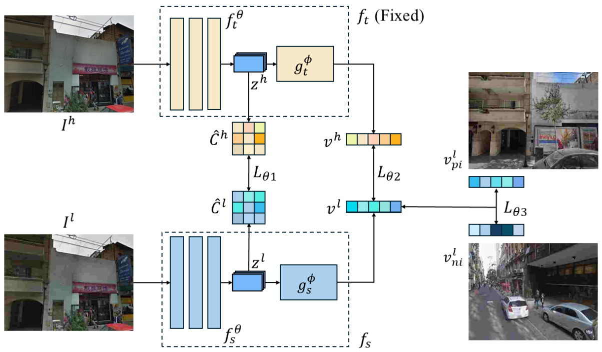

Figure 2. **Proposed Knowledge Distillation Methodology:** Through ICKD loss between latent codes z^h and z^l, MSE loss between global descriptors v^h and v^l, and triplet loss on v^l, the student branch f_s learns the teacher branch f_t's knowledge of high quality images I^h to extract more discriminative, representative v^l from low quality images I^l.

After knowledge distillation was first proposed in [13], it has been applied to various computer-vision tasks such as super-resolution [15], image classification [22, 36], and segmentation [22]. More recently, knowledge distillation has also been used specifically in VPR to produce more discriminative global descriptors, such as learning from structural knowledge of an image-segmentation model [27]. We believe that knowledge distillation could similarly improve VPR accuracy for low-quality images.

3. Distillation Methodology

Our goal is to enhance existing VPR methods by enabling them to extract more representative global descriptors from low-quality images. To achieve this, we apply knowledge distillation techniques directly at the descriptor extraction level. Specifically, we enable a student descriptor extractor, which processes low-quality images, to approximate the outputs of a teacher network operating on high-quality images. This approach leverages the rich, discriminative information from high-quality images to compensate for the information loss in degraded inputs.

3.1. Conceptual Model of Distillation for VPR

Given some VPR method, we initialize two copies, designating one as the teacher branch f_t and the other as the student branch f_s. The teacher processes a high-quality image I^h while the student processes its low-quality counterpart I^l. As illustrated in Fig. 2, the teacher model comprises a feature encoder f_t^θ and a descriptor aggregator g_t^ϕ. Similarly, the student consists of a structurally identical encoder f_s^θ and aggregator g_s^ϕ. Between the encoder and aggregator, the teacher and student produce intermediate 3D latent codes z^h and z^l, respectively. These latent codes are subsequently aggregated into global descriptors v^h and v^l. Our distillation strategy is built on three distinct loss functions designed to transfer the high-quality feature representation from the teacher to the student.

3.2. ICKD Loss

The Inter-Channel Correlation Knowledge Distillation (ICKD) loss [22] is employed to ensure that the student network's latent code z^l closely mimics the teacher network's latent code z^h in a channel-wise manner. In Visual Place Recognition (VPR), each channel of the latent code can be viewed as a local feature descriptor, and it is critical that these features remain consistent regardless of image quality. By enforcing similarity in the inter-channel correlations between z^l and z^h, we help preserve the discriminative structural information necessary for robust localization.

Owing to potential differences in spatial dimensions between I^l and I^h, the latent codes z^l and z^h may also differ in size, making traditional distillation losses inapplicable. To overcome this, the ICKD loss computes an Inter-Channel Correlation (ICC) matrix from the latent codes, effectively reducing their spatial dimensions to a consistent $c \times c$ form.

More specifically, let I^l have resolution $W \times H$ and let the latent codes z^h and z^l have dimensions $c \times W' \times H'$ and $c \times w' \times h'$, respectively. By concatenating the spatial dimensions w' and h' for z^l, we form a vector p^l of size $c \times (w' \cdot h')$. This vector is normalized along its second dimension to yield $\hat{p^l} = \frac{p^l}{\|p^l\|}$. The ICC matrix is then computed as $C^l = \hat{p^l}\hat{p^l}^T$ (with dimensions $c \times c$) and normalized to obtain $\hat{C^l} = \frac{C^l}{\|C^l\|}$. A similar process applied to z^h results in $\hat{C^h}$. The ICKD loss is formulated as:

$$L_{\theta 1} = \|\hat{C^l} - \hat{C^h}\|_2. \tag{1}$$

ensuring that the student encoder learns to mimic the channel correlations of the teacher, thereby preserving discriminative structural information despite spatial resolution differences.

3.3. Mean Squared Error (MSE) Loss

To directly align the global descriptors produced by the teacher and student branches, we employ a Mean Squared Error (MSE) loss. This loss minimizes the difference between v^l and v^h, enforcing that the student produces descriptors similar to those of the teacher despite the quality degradation of its input:

$$L_{\theta 2} = \|v^l - v^h\|_2^2. \tag{2}$$

This term provides a straightforward supervision signal, encouraging the student to closely emulate the teacher's output.

3.4. Weakly Supervised Triplet Ranking Loss

The weakly supervised Triplet Ranking Loss is applied solely within the student branch f_s to further enhance the discriminative power of the learned descriptors. In our framework, following the same strategy as NetVLAD [3], positive and negative samples are defined based on geographic proximity: images captured within a predefined distance (e.g., 25 meters) of the query are treated as positives, while those farther away are considered negatives. Inspired by this framework, our loss ensures that the descriptor v^l of a query image is closer (in Euclidean distance) to that of a geographically proximate (positive) image v_p^l than to that of a distant (negative) image v_n^l. Formally, the loss enforces:

$$d_\theta(v^l, v_p^l) < d_\theta(v^l, v_n^l),$$

where d_θ denotes the Euclidean distance.

The triplet loss is expressed as:

$$L_{\theta 3} = \sum_j l\left(\min_i d_\theta^2(v^l, v_{p_i}^l) - d_\theta^2(v^l, v_{nj}^l) + m\right), \quad (3)$$

with $l(x) = \max(x, 0)$ being the hinge loss and m a margin parameter that ensures a sufficient gap between positive and negative pairs. Here, the term $\min_i d_\theta^2(v^l, v_{p_i}^l)$ selects, among a set of candidate positives in the batch, the descriptor that is closest to the query. This selection is performed based on the current descriptor distances, which serves as a weak supervision signal because it relies on geographic proximity and descriptor similarity rather than explicit ground-truth labels. Although this strategy may occasionally select a false positive, the margin m and continuous network updates help mitigate this risk by ensuring that the model learns to focus on the most challenging and truly similar examples.

3.5. Composite Loss

Finally, the overall distillation objective is defined as a composite loss combining the three individual terms:

$$L = L_{\theta 1} + \alpha L_{\theta 2} + \beta L_{\theta 3}, \quad (4)$$

where α and β are weighting coefficients that balance the influence of the MSE and Triplet losses relative to the ICKD loss. In our experiments, we analyze both the individual and combined effects of these losses to validate the efficacy of our distillation approach.

4. Experiments

To evaluate the effectiveness of our distillation methodology, we analyze the impact of each loss component ($L_{\theta 1}$, $L_{\theta 2}$, $L_{\theta 3}$) as well as their composite form L on VPR performance. We perform distillation on the student branch f_s using the GSV-Cities VPR dataset [1] and subsequently evaluate retrieval accuracy on several benchmark datasets from the Deep Visual Geo-localization Benchmark [6]. For all experiments, low-quality images I^l are generated by applying 90% JPEG compression (quality level = 10) to the original high-quality images I^h. This choice allows us to simulate bandwidth-constrained scenarios while keeping the degradation consistent across training and testing.

Following the protocol in [33], we evaluate f_s using a database and query set that both consist of I^l images. This controlled setting avoids inconsistencies due to variable compression levels and also illustrates the reduced cost of database construction when operating under strict quality constraints. Although our current experiments focus on JPEG compression, we further explore the generalizability of our method to low image resolution and high video quantization in Section 6.

4.1. Datasets Configurations

4.1.1. Training Dataset

We use the GSV-Cities dataset, which contains approximately 530,000 street view images from 23 cities. To reduce training cost and assess our method under constrained data conditions, we use only the low-resolution subset (400×300), reducing the training set to about 194,000 images.

4.1.2. Testing Datasets

We evaluate f_s on three standard VPR datasets: Mapillary SLS[38], Nordland[30], and Tokyo 24/7[35]

4.2. VPR Methods

To validate the broad applicability of our knowledge distillation framework, we experiment with five VPR methods representing different architectures:

- NetVLAD [3] and MixVPR [2] (CNN-based),
- AnyLoc [17], DINOv2SALAD [16], and CricaVPR [23] (DINOv2-based [25]).

For each method, we initialize both the teacher f_t and student f_s using pretrained weights released by each method's authors for faster convergence. To investigate the effects of individual loss components, we train f_s under all seven possible combinations of $L_{\theta 1}$, $L_{\theta 2}$, and $L_{\theta 3}$ (with at least one term used from Eq. (4)).

As a baseline, we also fine-tune each method (except AnyLoc, which uses a frozen DINOv2 backbone) on the same I^l images using pretrained weights and the authors' provided training procedure instead of our distillation losses. This baseline comparison highlights the improvement attributable solely to our distillation framework.

4.3. Experimental Setup

4.3.1. Training Configuration

Our loss weighting coefficients in Eq. (4) are set to $\alpha = 10^5$ and $\beta = 10^4$. These values were chosen based on preliminary ablation studies aimed at balancing the scale of the

Table 1. **Best Performance of VPR Methods after Distillation:** For low quality images, the loss combination producing the highest recall rate is compared against the baseline of fine-tuning. For both distilled and fine-tuned weights, the change in VPR recall is represented as a delta relative to the performance of each method using pretrained weights. Within each dataset and method, greene text indicates the greatest improvement for every R@N, whereas any decrease relative to pretrained performance is marked as red. The recall rates using pretrained weights on unmodified I^h images are provided as reference.

VPR Methods	Configuration	Mapillary SLS				Nordland				Tokyo 24/7			
		R@1	R@2	R@5	R@10	R@1	R@2	R@5	R@10	R@1	R@2	R@5	R@10
MixVPR	pretrained (I^h)	82.73	86.67	89.73	91.65	57.79	64.13	71.49	76.41	87.30	89.52	92.06	93.65
	pretrained (I^l)	71.87	76.61	81.22	84.19	31.05	36.12	44.13	51.23	66.03	73.33	78.73	82.22
	finetuned	+4.43	+3.86	+3.60	+2.99	+13.37	+14.28	+14.64	+14.46	+9.52	+7.62	+6.03	+5.40
	ICKD	**+4.33**	**+3.69**	**+3.20**	**+2.93**	**+15.00**	**+16.59**	**+16.78**	**+15.94**	**+8.25**	**+6.03**	**+6.35**	**+6.67**
CricaVPR	pretrained (I^h)	74.74	80.92	86.09	88.75	87.64	90.65	94.24	95.80	90.16	92.38	95.56	96.19
	pretrained (I^l)	68.14	74.41	80.24	83.19	63.51	69.71	77.46	82.64	74.29	79.68	84.76	87.94
	finetuned	-0.12	-0.64	-0.77	-0.23	-2.97	-2.93	-2.97	-2.79	+5.08	+4.44	+2.54	+1.90
	ICKD	**+0.89**	**+1.39**	**+1.11**	**+1.14**	**+11.63**	**+10.80**	**+8.95**	**+7.03**	**+6.98**	**+5.08**	**+6.03**	**+4.76**
DINOv2 SALAD	pretrained (I^h)	89.20	92.40	94.70	95.84	88.08	90.94	94.13	95.98	97.14	97.46	98.73	99.05
	pretrained (I^l)	84.60	88.85	91.89	93.68	67.90	73.88	80.40	84.24	89.21	92.06	95.87	96.51
	finetuned	-0.22	-0.57	-0.60	-0.76	+0.04	-0.40	-0.76	-0.04	-0.63	0.00	-0.95	0.00
	ICKD	**+0.52**	**+0.45**	**+0.20**	**+0.30**	**+1.56**	**+1.81**	**+1.09**	**+1.20**	**+1.59**	**+1.27**	**+0.32**	0.00
NetVLAD	pretrained (I^h)	49.29	55.24	62.07	67.29	5.51	6.67	8.62	11.38	60.63	63.81	69.21	74.29
	pretrained (I^l)	32.60	37.67	44.92	50.13	1.99	3.01	4.93	7.07	27.94	33.02	41.90	47.94
	finetuned	+0.03	+0.01	-0.02	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	MSE	**+4.46**	**+5.32**	**+5.09**	**+5.01**	**+0.54**	**+0.43**	-0.18	-0.72	**+8.25**	**+6.35**	**+6.03**	**+5.08**
AnyLoc	pretrained (I^h)	56.51	62.31	68.28	72.89	12.90	16.23	20.18	24.28	88.25	91.11	94.92	96.83
	pretrained (I^l)	48.04	56.45	63.84	69.17	10.11	12.97	17.86	21.92	83.49	87.30	91.75	95.87
	ICKD + Triplet	**+0.99**	**+1.54**	**+2.68**	**+2.44**	**+1.74**	**+1.88**	**+1.52**	**+2.54**	-5.71	-1.59	**+1.59**	-0.95

MSE and triplet losses relative to the ICKD loss. The learning rate is initialized at 10^{-5} and adjusted using a decay factor of 2×10^{-11} along with an exponential decay rate of 0.99999. Each training experiment (one of the seven loss combinations applied to one of the five VPR methods) comprises a single epoch over the reduced GSV-Cities dataset (Section 4.1.1). For the triplet loss (Eq. (3)), we sample 5 negative images per training sample to form v_n^l.

4.3.2. VPR Performance Evaluation

We assess the VPR performance of the distilled student network f_s using the Recall at N (R@N) metric, following prior work [2, 3, 5, 12, 16, 17, 23]. R@N is defined as the percentage of queries for which at least one of the top N retrieved database images is within a specified distance threshold d of the query's true location. Following the VPR methods we experimented with, we set d to 25 meters, and we evaluate R@N for N values of 1, 2, 5, and 10.

5. Results and Discussions

We analyze the performance of the student network f_s after distillation using each loss combination. We report quantitative improvements in VPR recall rates, discuss observed trends across different methods and datasets, and provide qualitative visualizations to interpret the behavior of the network after distillation.

5.1. Quantitative Results

Table 1 demonstrates the applicability of (4) and its components to a representative selection of VPR methods and datasets. Albeit with varying magnitudes, the deltas of recall rates under v^l after distillation and those under v^l from pretrained weights are generally positive. However, fine-tuning is inconsistent and mostly less effective than distillation. The Nordland recall rates for MixVPR and CricaVPR, two architecturally distinct methods, exemplifies distillation's superior efficacy and applicability. There are only two instances where distillation partially failed to increase VPR recall rates under v^l for some method and dataset, namely NetVLAD on Nordland and AnyLoc on Tokyo 24/7. As AnyLoc's recall rates fluctuate greatly on Tokyo 24/7, we postulate the combination of the natures of this method and dataset to be an edge case. Since NetVLAD's improvement on Nordland for all R@N is less than that on other datasets, we reserve this phenomenon for further analysis in 5.2.

Besides numerically demonstrating the most effective distillation losses, we graphically compare each loss combination in fig. 3, where we analyze full recall rate data for all variants of f_s. Previously, we have observed that ICKD (1) produces best VPR performances for MixVPR and CricaVPR and that MSE (2) is best for NetVLAD, and these three methods benefit the most from distillation. Now, the relative superiority of the two aforementioned losses could be seen on all methods, where the recall rates for

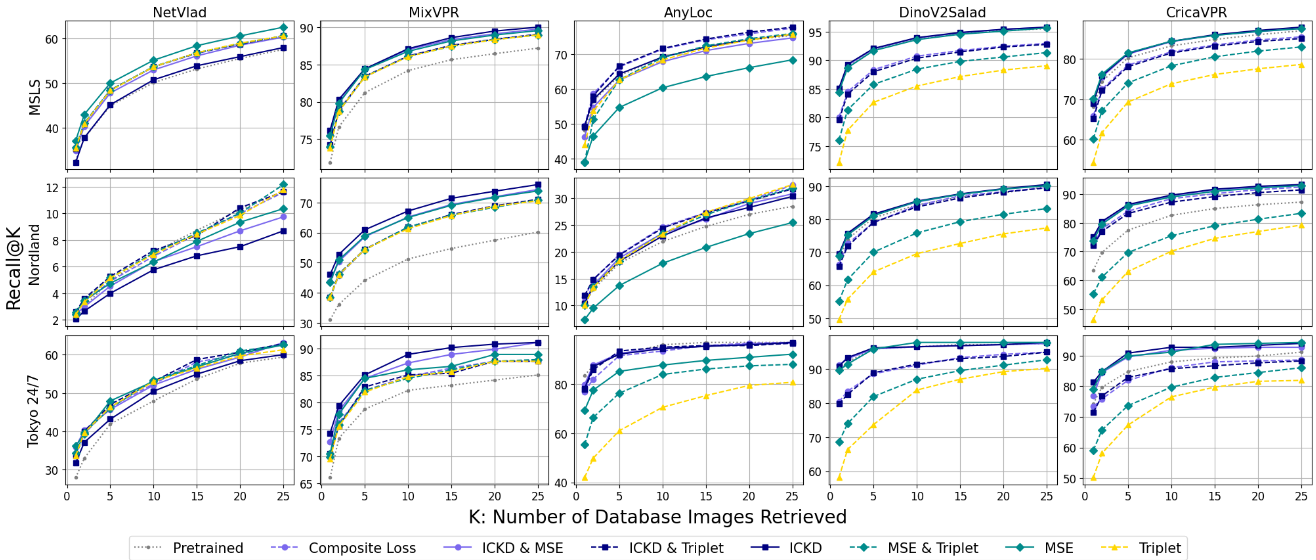

Figure 3. **Recall Rates for All Losses:** For each method, we compare f_s after distillation with loss combinations as well as pretrained weights. Different marker shapes indicate different combinations of ICKD and MSE losses, whereas including the triplet loss is shown with a dashed line. For readability, the y-axis of each subplot is independently scaled.

Table 2. Indoor Evaluation Using Pretrained Weights

Methods	Quality	TUM LSI		Gangnam Station		NYC-Indoor-VPR	
		R@1	R@5	R@1	R@5	R@1	R@5
MixVPR	I^h	94.09	99.55	4.39	13.63	41.23	83.76
	90%	91.36	99.09	3.93	11.76	40.04	80.76
CricaVPR	I^h	92.27	99.09	7.48	21.64	37.92	82.36
	90%	92.27	97.27	7.67	21.37	37.30	79.26
DINOv2 SALAD	I^h	94.09	99.09	8.93	22.67	39.89	82.98
	90%	91.82	98.64	9.43	21.37	39.58	81.69
NetVLAD	I^h	95.00	99.09	3.21	8.36	39.42	81.27
	90%	91.82	98.18	2.94	9.81	38.23	79.10
AnyLoc	I^h	97.73	99.55	4.47	12.79	37.92	82.77
	90%	93.64	98.18	3.63	11.11	36.37	81.07

f_s trained under either or both of them generally trend towards the top of all loss combinations. Here, AnyLoc and NetVLAD on Nordland are exceptions, where augmenting (1) or (2) with triplet loss (3) produces a better-performing f_s. However, besides NetVLAD, whose original training scheme uses the triplet loss, (3) is generally the least effective when used alone.

5.2. Qualitative results

We further investigate distillation's effects on MixVPR, CricaVPR, and NetVLAD, the most improved methods, and potentially account for the latter's more limited improvement on Nordland. In fig. 4, we plot activation maps for each method's encoder f_s^θ after distillation with their respective best-performing loss. From each of the three VPR testing datasets in 4.1.2, we visualize an I^l query where f_s improves upon f_t. On Mapillary SLS and Nordland, distillation mitigates the distraction of uninformative sky features especially for NetVLAD and MixVPR. More generally, distillation shifts focus from repetitive foreground features, such as road surfaces and ground-level vegetation, to more structurally distinct whole-scene features, such as the contours of buildings and farther tree lines. In the case of NetVLAD, the highly selective focus of f_s^θ could account for its lesser recall rate improvement on Nordland. For the other two methods, while MixVPR's improvement could be more intuitively attributed to increased overall focus on informative image regions, CricaVPR's comparatively smaller change in focus suggests that its improvements originates from other factors, possibly concerning its aggregator g_s^ϕ.

6. Additional Experiments

While our knowledge distillation approach already demonstrates notable efficacy for the majority of methods and datasets tested, we extend distillation to more datasets and other modes of image quality reduction discussed in 2.1.

6.1. JPEG Compression's Impacts on Indoor Datasets

While our previous VPR testing datasets cover both rural and urban environments and contain temporal changes and perceptual aliasing, they all capture outdoor environments. In table 2, We further examine the impacts of JPEG compression on three indoor datasets of different dif-

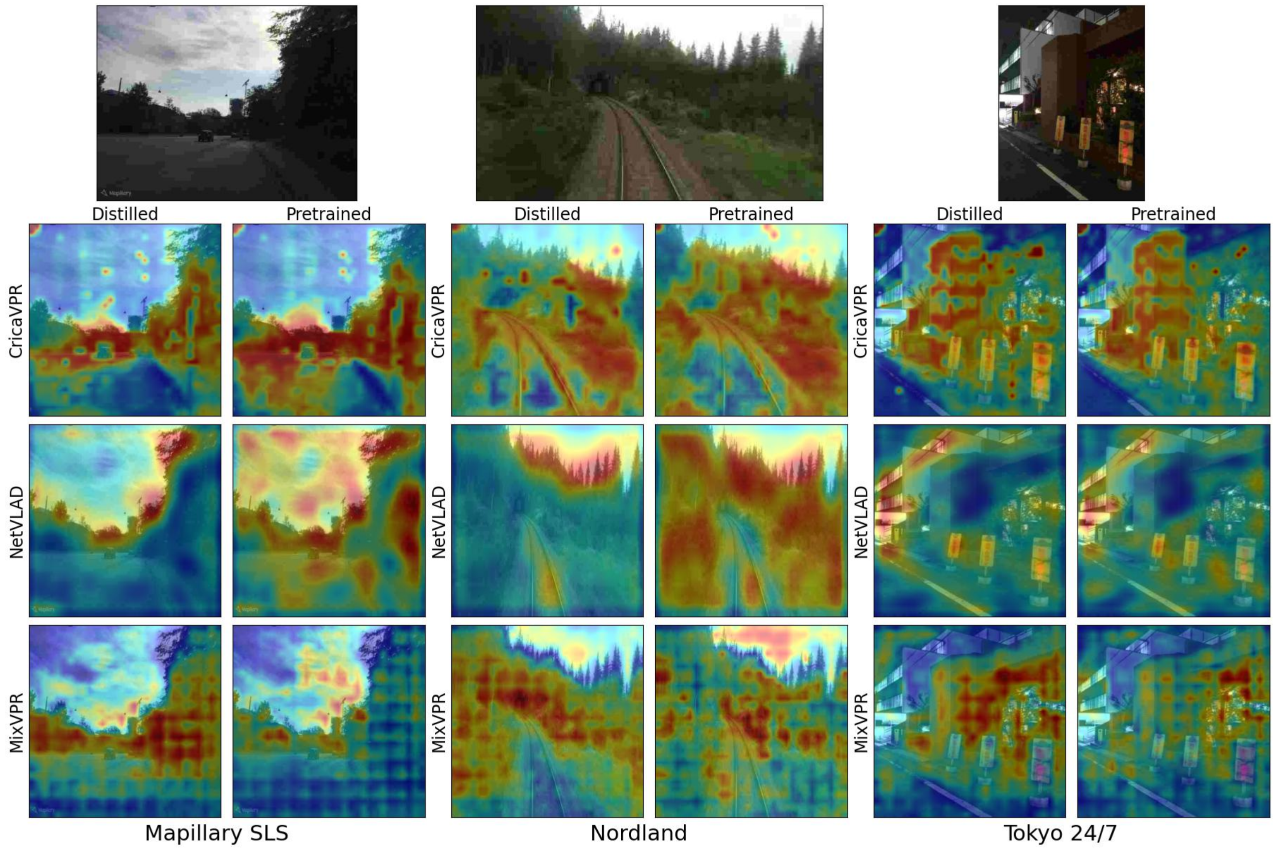

Figure 4. **Activation Maps:** The selected I^l queries (top of figure) are successfully recalled by f_s, while f_t fails. For CricaVPR, the focus of its ICKD-trained f_s is visualized by averaging its z^l across the channel dimension in accordance with its authors. For NetVLAD, we averaged a weighted sum of its MSE-trained z^l via tentative cluster assignments from g_s^{ϕ}. For MixVPR, since its authors did not visualize activations, we masked small regions of I^l and considered the resulting change in its ICKD-trained v_l as activations, creating block-like patches on its heatmap.

ficulty levels, namely TUM LSI [37], Gangnam Station (one of NAVER LABS' large-scale localization datasets [21]), and NYC-Indoor-VPR [29]. While the performance of all methods with pretrained weights fluctuate greatly across datasets, JPEG compression's impacts on recall rates are significantly smaller than previous results with outdoor datasets. We postulate that due to indoor environments generally lack large contiguous regions of uninformative features such as sky, the proportion of distinctive features degraded by JPEG compression is smaller within indoor images. Therefore, instead of repeating our previous experiments, we extend our distillation methodology to other possible forms of I^l.

6.2. Other Modalities of Image Quality Reduction

As JPEG compression had greater impacts on outdoor scenes, we use outdoor datasets from 4.1.2 to yield I^l with reduced resolution instead of JPEG compression. Secondly, we explore the influence of video quantization on VPR recall, mirroring realistic practice of streaming image data through video. However, the image-based nature of existing VPR datasets prompts us to curate a custom video-based VPR dataset to produce I^l with increased video quantization.

Among the five methods in table 1, NetVLAD suffers the most from lowered image quality in general. Thus, we select NetVLAD as the candidate of verifying knowledge distillation's benefits for VPR performance under resolution and quantization-based I^l.

6.3. Custom Video-Based VPR Dataset

Our custom VPR dataset is sourced from indoor spaces of the 6th floor of the Lighthouse Guild, an eye care facility in New York City. Instead of following existing indoor VPR datasets in capturing images [21, 28, 29, 37], we record panoramic videos using the Insta360 camera, providing a 360° view. Each frame, with dimensions 5760×2880, was horizontally segmented into 18 perspective images with 90° field of view and dimensions 1440×810. These images were assembled into perspective videos for both resolution

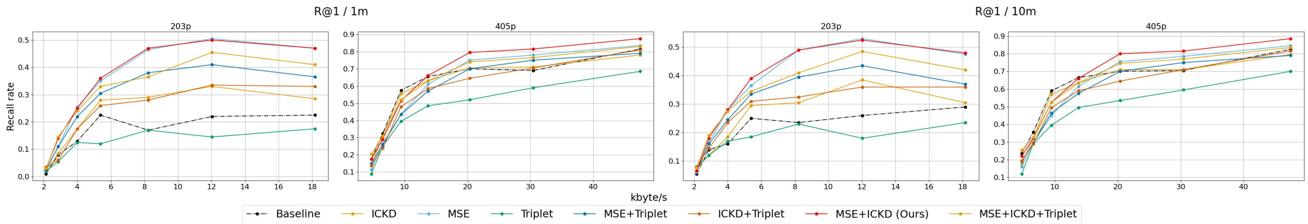

Figure 5. **Results on Our Dataset:** The recall rates of Pitts250k-trained f_s are obtained for each quantization parameter. For ease of visualization, the latter is converted into the video bitrate metric with lower bitrate indicating higher quantization. The left two plots were calculated with a VPR threshold of 1m, and the right two have a 10m threshold.

reduction and video quantization according to the following:

- Resolutions: 405p (720×405), 203p (360×203)
- Quantization Parameters: 30, 33, 36, 39, 42, 45, 48

6.4. Resolution and Quantization Experimental Results

We not only validate distillation's efficacy for different forms of I^l but also its independence to datasets and I^l's modality during training. While following 4.3.1's steps, we perform distillation on NetVLAD using the Pitts250k dataset [34], which contains around $250,000$ outdoor 640×480 images of Pittsburgh's urban environments. All images are reduced to 240×180 as I^l for f_s to learn on. To analyze distillation's robustness to changing I^l modalities, we test NetVLAD's recall on our custom dataset with v^l extracted from resolutions of 405p and 203p and quantization parameters spanning from 30 to 48 in intervals of 3. For previous outdoor datasets, we extract v^l from 320×240 (240p). The influence of our dataset's various resolutions and quantization levels on f_s's recall rate after distillation is compared against pretrained weights as a baseline in Fig. 5. The 203p results restate MSE (2)'s efficacy on NetVLAD, but despite ICKD (1)'s comparatively inferior performance on constant-modality I^l before, combining the two losses further improves VPR performance on 405p, suggesting ICKD's greater adaptability under video quantization. Notably, changing recall distance thresholds as defined in 4.3.2 has little effect on VPR performance.

For outdoor datasets, we compare pretrained NetVLAD against trained f_s with the highest recall rate in table 3. NetVLAD's best-performing distillation loss remains MSE, but adding ICKD yields stronger performance on Tokyo 24/7. This corroborates the previous observation of ICKD's adaptability, which may allow f_s to better withstand changing I^l modalities from 180p during distillation to 240p in testing.

7. Conclusion

Addressing the concern of reduced VPR accuracy under the realistic scenario of image quality reduction, our knowledge distillation methodology for extracting more discriminative descriptors from low quality images achieves significant performance gains under JPEG and video compression for various VPR methods and datasets. Furthermore, we have demonstrated our approach's potential to generalize to other kinds of low quality images such as low resolutions, reducing the literature gap on video-based VPR datasets in the process.

A suitable direction of future work is to justify distillation's effects on different kinds of VPR methods. While we have identified the general strengths of ICKD and MSE losses and noted the specific usability of the triplet loss, a better understanding of each VPR method's affinity to specific losses could reveal further insights on VPR methods themselves. Secondly, more VPR datasets and research under the context of video quantization could prove valuable for the future real-world applications of VPR.

Table 3. NetVLAD's Performance on 240p Images

Dataset	Configuration	R@1	R@2	R@5	R@10
Mapillary SLS	pretrained	31.16	37.89	46.36	52.54
	MSE	**+7.57**	**+7.88**	**+7.64**	**+7.54**
Nordland	pretrained	3.99	5.40	8.34	10.40
	MSE	**+1.12**	**+1.23**	**+1.23**	**+1.81**
Tokyo 24/7	pretrained	24.84	29.62	39.49	48.41
	ICKD + MSE	**+7.64**	**+8.28**	**+7.64**	**+7.32**

Acknowledgment

We would like to express our gratitude to Zezheng Li and Liyuan Geng for their valuable assistance with test data preprocessing, accelerating this work's progress.

References

[1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022. 2, 4

[2] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Mixvpr: Feature mixing for visual place recognition. In *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, pages 2998–3007, 2023. 1, 4, 5

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.,*, pages 5297–5307, 2016. 2, 4, 5

[4] Bruno Arcanjo, Bruno Ferrarini, Michael Milford, Klaus D. McDonald-Maier, and Shoaib Ehsan. An efficient and scalable collection of fly-inspired voting units for visual place recognition in changing environments. *IEEE Robot. Automat. Lett.*, 7(2):2527–2534, 2022. 2

[5] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 4878–4888, 2022. 2, 5

[6] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022. 4

[7] Helen Carson, Jason J. Ford, and Michael Milford. Unsupervised quality prediction for improved single-frame and weighted sequential visual place recognition. In *Proc. IEEE Int. Conf. Robot. Automat.*, pages 3955–3961, 2023. 2

[8] Anh-Dzung Doan, Yasir Latif, Tat-Jun Chin, Yu Liu, Thanh-Toan Do, and Ian Reid. Scalable place recognition under appearance change for autonomous driving. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 9319–9328, 2019. 1

[9] Bruno Ferrarini, Michael J. Milford, Klaus D. McDonald-Maier, and Shoaib Ehsan. Binary neural networks for memory-efficient and effective visual place recognition in changing environments. *IEEE Trans. Robot.*, 38(4):2617–2631, 2022. 2

[10] Sourav Garg and Michael Milford. Fast, compact and highly scalable visual place recognition through sequence-based matching of overloaded representations. In *Proc. IEEE Int. Conf. Robot. Automat.*, pages 3341–3348. IEEE, 2020. 1, 2

[11] Sourav Garg, Tobias Fischer, and Michael Milford. Where is your place, visual place recognition? In *Proc. Int. Jt. Conf. Artif. Intell.*, pages 4416–4425, 2021. 1

[12] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 14141–14152, 2021. 5

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 3

[14] Graham Hudson, Alain Léger, Birger Niss, Istvan Sebestyen, and Jørgen Vaaben. Jpeg-1 standard 25 years: Past, present, and future reasons for a success. *Journal of Electronic Imaging*, 27:1, 2018. 2

[15] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.,*, pages 723–731, 2018. 3

[16] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024. 4, 5

[17] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robot. Automat. Lett.*, 2023. 4, 5

[18] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023. 2

[19] Anastasiia Kornilova, Ivan Moskalenko, Timofei Pushkin, Fakhriddin Tojiboev, Rahim Tariverdizadeh, and Gonzalo Ferrer. Dominating set database selection for visual place recognition. In *Proc. 21st Int. Conf. Adv. Robot.*, pages 473–479. IEEE, 2023. 2

[20] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018. 2

[21] Donghwan Lee, Soohyun Ryu, Suyong Yeon, Yonghan Lee, Deokhwa Kim, Cheolho Han, Yohann Cabon, Philippe Weinzaepfel, Nicolas Guérin, Gabriela Csurka, et al. Large-scale localization datasets in crowded indoor spaces. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 3227–3236, 2021. 7

[22] Li Liu, Qingle Huang, Sihao Lin, Hongwei Xie, Bing Wang, Xiaojun Chang, and Xiaodan Liang. Exploring inter-channel correlation for diversity-preserved knowledge distillation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 8271–8280, 2021. 3

[23] Feng Lu, Xiangyuan Lan, Lijun Zhang, Dongmei Jiang, Yaowei Wang, and Chun Yuan. Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024. 1, 4, 5

[24] Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547, 2021. 1, 2

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V., Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 4

[26] JoonKyu Park, Sanghyun Son, and Kyoung Mu Lee. Content-aware local gan for photo-realistic super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10585–10594, 2023. 2

[27] Yanqing Shen, Sanping Zhou, Jingwen Fu, Ruotong Wang, Shitao Chen, and Nanning Zheng. Structvpr: Distill structural knowledge with weighting samples for visual place recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 11217–11226, 2023. 3

[28] Diwei Sheng, Yuxiang Chai, Xinru Li, Chen Feng, Jianzhe Lin, Claudio Silva, and John-Ross Rizzo. Nyu-vpr: long-term visual place recognition benchmark with view direction and data anonymization influences. In *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 9773–9779. IEEE, 2021. 7

[29] Diwei Sheng, Anbang Yang, John-Ross Rizzo, and Chen Feng. Nyc-indoor-vpr: A long-term indoor visual place

recognition dataset with semi-automatic annotation. In *Proc. IEEE Int. Conf. Robot. Automat.*, pages 14853–14859, 2024. 7

[30] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. *Proc. of Workshop on Long-Term Autonomy, IEEE Int. Conf. Robot. Automat.*, page 2013, 2013. 2, 4

[31] Mihnea-Alexandru Tomita, Bruno Ferrarini, Michael Milford, Klaus McDonald-Maier, and Shoaib Ehsan. Data efficient visual place recognition using extremely jpeg-compressed images, 2023. 2

[32] Mihnea-Alexandru Tomita, Bruno Ferrarini, Michael Milford, Klaus McDonald-Maier, and Shoaib Ehsan. Visual place recognition with low-resolution images, 2023. 2

[33] Mihnea-Alexandru Tomita, Bruno Ferrarini, Michael Milford, Klaus Dieter Mcdonald-Maier, and Shoaib Ehsan. Data-efficient sequence-based visual place recognition with highly compressed jpeg images. *ArXiv*, abs/2302.13314, 2023. 2, 4

[34] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.,*, pages 883–890, 2013. 8

[35] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(2):257–271, 2018. 2, 4

[36] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 1365–1374, 2019. 3

[37] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017. 7

[38] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020. 2, 4

[39] Zhe Xin, Yinghao Cai, Tao Lu, Xiaoxia Xing, Shaojun Cai, Jixiang Zhang, Yiping Yang, and Yanqing Wang. Localizing discriminative visual landmarks for place recognition. In *Proc. IEEE Int. Conf. Robot. Automat.*, pages 5979–5985. IEEE, 2019. 1

[40] Anbang Yang, Mahya Beheshti, Todd E Hudson, Rajesh Vedanthan, Wachara Riewpaiboon, Pattanasak Mongkolwat, Chen Feng, and John-Ross Rizzo. Unav: An infrastructure-independent vision-based navigation system for people with blindness and low vision. *Sensors*, 22(22):8894, 2022. 2

[41] Peng Yin, Lingyun Xu, Xueqian Li, Chen Yin, Yingli Li, Rangaprasad Arun Srivatsan, Lu Li, Jianmin Ji, and Yuqing He. A multi-domain feature learning method for visual place recognition. In *Proc. IEEE Int. Conf. Robot. Automat.*, pages 319–324. IEEE, 2019. 1

[42] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36:13294–13307, 2023. 2