Through the Lens of Doubt: Robust and Efficient Uncertainty Estimation for Visual Place Recognition

Emily Miller[1], Michael Milford[2], Muhammad Burhan Hafez[1], Sarvapali Ramchurn[1], Shoaib Ehsan[1,3]

***Abstract*—Visual Place Recognition (VPR) enables robots and autonomous vehicles to identify previously visited locations by matching current observations against a database of known places. However, VPR systems face significant challenges when deployed across varying visual environments, lighting conditions, seasonal changes, and viewpoints changes. Failure-critical VPR applications, such as loop closure detection in simultaneous localization and mapping (SLAM) pipelines, require robust estimation of place matching uncertainty. We propose three training-free uncertainty metrics that estimate prediction confidence by analyzing inherent statistical patterns in similarity scores from any existing VPR method. Similarity Distribution (SD) quantifies match distinctiveness by measuring score separation between candidates; Ratio Spread (RS) evaluates competitive ambiguity among top-scoring locations; and Statistical Uncertainty (SU) is a combination of SD and RS that provides a unified metric that generalizes across datasets and VPR methods without requiring validation data to select the optimal metric. All three metrics operate without additional model training, architectural modifications, or computationally expensive geometric verification. Comprehensive evaluation across nine state-of-the-art VPR methods and six benchmark datasets confirms that our metrics excel at discriminating between correct and incorrect VPR matches, and consistently outperform existing approaches while maintaining negligible computational overhead, making it deployable for real-time robotic applications across varied environmental conditions with improved precision-recall performance.**

***Index Terms*—Localization, Vision-Based Navigation, Deep Learning for Visual Perception**

I. Introduction

VISUAL Place Recognition (VPR) is a core component of robotic localization and autonomous navigation systems [1], enabling robots to localize by matching current observations with georeferenced image databases [2]. Despite advances in descriptor-based methods [3], reliable deployment in failure-critical applications such as autonomous driving remains challenging due to limited uncertainty quantification. Mission-critical tasks like loop closure detection in SLAM pipelines [4] require robust confidence estimates to avoid failures caused by perceptual aliasing, seasonal variation, and repetitive structures in changing environments and viewpoints.

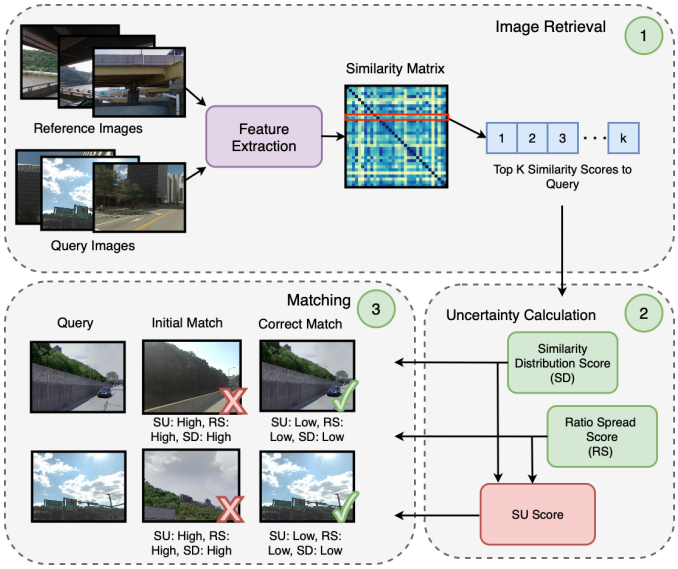

Fig. 1: Overview of our training-free uncertainty metrics integrated into the VPR pipeline, operating directly on similarity scores to provide confidence estimates without model modifications. In Stage 3 of the diagram, the initial incorrect match shows high uncertainty, while the subsequent correct match has the lowest uncertainty among the top-k on the Pittsburgh250k dataset using Eigenplaces.

Current uncertainty estimation methods face practical limitations that prevent their use in failure-intolerant applications. Retrieval-based approaches such as L2 distance [5] and Perceptual Aliasing Score [6] show inconsistent performance across datasets and environments, with metric selection depending on the deployment context. Deep Uncertainty Estimation methods like Bayesian Triplet Loss [7] and STUN [8] require retraining for each new VPR architecture [4], while Geometric Verification methods using SIFT [9], DELF [10], and SuperPoint [11] are computationally prohibitive for real-time operation. Selecting the most reliable metric typically requires validation data that may be unavailable in unseen environments, posing risks in safety-critical systems where incorrect matches can lead to catastrophic failures [12].

This work introduces training-free uncertainty metrics that operate directly on VPR similarity scores without requiring retraining or architectural changes. The metrics are derived from the distribution of top-k matches: Similarity Distribution (SD) measures match distinctiveness through the gap between

Manuscript received: November, 4, 2025; Revised February, 2, 2026; Accepted March, 3, 2026.

This paper was recommended for publication by Editor Pascal Vasseur upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the U.K. Engineering and Physical Sciences Research Council under Grant EP/Y009800/1 and Grant EP/V00784X/1.

[1]Emily Miller, Muhammad Burhan Hafez and Sarvapali Ramchurn are with School of Electronics and Computer Science, University of Southampton, SO17 1BJ Southampton, U.K. (e-mail: em3g20@soton.ac.uk, burhan.hafez@soton.ac.uk, sdr1@soton.ac.uk).

[2]Michael Milford is with School of Electrical Engineering and Robotics, QUT Centre for Robotics, Brisbane, QLD 4000, Australia. (e-mail: michael.milford@qut.edu.au).

[1,3]Shoaib Ehsan is with the School of Electronics and Computer Science, University of Southampton, SO17 1BJ Southampton, U.K., and also with the School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Colchester, U.K. (e-mail: s.ehsan@soton.ac.uk).

Digital Object Identifier (DOI): see top of this page.

the highest and median similarity scores, while Ratio Spread (RS) captures ambiguity by analyzing score clustering among top candidates. We further propose Statistical Uncertainty (SU), which combines SD and RS to provide a validation-free metric that generalizes across VPR methods and datasets while maintaining the efficiency required for real-time operation.

Our main contributions are:

- We introduce three training-free uncertainty metrics for VPR, Similarity Distribution, Ratio Spread, and SU Score, that generalize across architectures and datasets without validation data.
- We evaluate nine VPR methods on six benchmarks, demonstrating improved failure detection (AUC-PR) and improved precision-recall performance when used for prediction rejection.

The rest of the paper is organized as follows. Section II reviews related work in VPR and uncertainty estimation. Section III introduces three uncertainty estimation metrics. Section IV outlines experimental setup. Results and analysis are given in Section V. Ablation studies are provided in Section VI. Finally, Section VII draws conclusions.

II. Related Work

This section provides an overview of the related work.

A. Visual Place Recognition

Visual Place Recognition (VPR) has evolved from handcrafted local features such as SIFT [9] and probabilistic frameworks like FAB-MAP [13] to deep learning approaches that learn robust hierarchical representations [3]. Modern VPR architectures generate compact global descriptors using feature aggregation layers. NetVLAD [14] introduced differentiable aggregation with learnable cluster centers, while later methods such as MixVPR [15] and TransVPR [16] incorporate multi-scale feature mixing and Transformer-based self-attention. These models are typically trained using Siamese or Triplet metric learning frameworks [17], [18] on large-scale datasets such as GSV-Cities [19].

B. Uncertainty Estimation in Visual Place Recognition

Robust uncertainty estimation proves crucial for safety-critical applications, where high-confidence incorrect matches from VPR systems can lead to catastrophic failures such as corrupted maps in Simultaneous Localization and Mapping (SLAM) pipelines [1]. Several works have demonstrated that rejecting predictions based on confidence estimates can significantly improve accuracy and prevent such failures [20], motivating the development of specialized uncertainty quantification methods for VPR systems. These VPR-specific methods draw inspiration from the broader deep learning community, where uncertainty is typically categorized as aleatoric (inherent data noise) and epistemic (model ignorance due to limited training data) [21]. General techniques to capture this include Monte Carlo Dropout [22], which approximates Bayesian inference by running multiple forward passes with active dropout layers, and Deep Ensembles [23], which train multiple models to measure prediction variance. Other approaches modify the model's output to directly predict a probability distribution instead of a point estimate [24]. However, these methods often impose significant computational overhead or require architectural modifications and extensive retraining, making them impractical for seamless integration with pre-existing, real-time VPR pipelines [25]. Our work bypasses these limitations by operating directly on the output similarity scores of any VPR model.

VPR-specific uncertainty estimation methods address these challenges through four main categories. Retrieval-based Uncertainty Estimation (RUE) methods estimate uncertainty directly from similarity scores, such as L2 distance or Perceptual Aliasing (PA) score, which is computed as the ratio between first and second nearest neighbor distances [5], [6]. While computationally efficient and training-free, RUE methods suffer from overconfidence in perceptual aliasing scenarios [20].

Data-driven Uncertainty Estimation (DUE) methods learn to infer uncertainty from image content, typically modeling aleatoric uncertainty. Approaches like Bayesian Triplet Loss [7] and self-supervised methods such as STUN [8] exemplify this category but require specialized training and exhibit sensitivity to domain shifts across different environments.

Geometric Verification (GV) approaches assess spatial consistency between query and database images through local feature matching [9]. Methods employing local features like DELF [10] or SuperPoint [11] with RANSAC-based verification achieve high robustness but incur computational costs prohibitive for real-time applications.

Spatial Uncertainty Estimation (SUE) methods [20] infer uncertainty from the spatial distribution of retrieved poses. Though effective in dense mapping scenarios, SUE performance depends on database quality and spatial coverage. Additionally, SUE requires spatial pose information that may be unavailable at runtime.

Beyond VPR, similar principles for uncertainty estimation have been explored in object recognition [26], semantic segmentation [27], and aerial navigation [28], where predictive uncertainty guides decision confidence and safety assessment.

III. Proposed Uncertainty Estimation Metrics

This section presents three uncertainty estimation metrics for VPR. Uncertainty in VPR arises from perceptual aliasing, the ratio between the first and second ranked similarity scores, where visually similar but geographically distinct locations compete strongly, and weak distinctiveness, where correct matches achieve only marginally higher scores due to environmental variations. Our approach estimates ranking confidence, the reliability of the top-ranked retrieval, by analyzing similarity score distributions post-hoc, without requiring specialized training or model modifications. Unlike Bayesian methods that model epistemic uncertainty through distributional outputs, we characterize whether the top candidate is sufficiently distinct from its competitors to be trusted.

A. Ratio Spread (RS)

It measures how strongly alternative candidates compete with the top-ranked match by evaluating how closely their

similarity scores approach the highest score. It is calculated as the average ratio of each subsequent candidate's score (from rank 2 to k) to the top score, reflecting the degree of competition among the most similar matches.

$$\text{RS} = \frac{1}{k-1} \sum_{i=2}^{k} \frac{s_i}{s_1} \tag{1}$$

where s_i is the similarity score of the i-th ranked place in descending order.

Higher RS values indicate that several candidates achieve similarity scores close to the top match, reflecting greater uncertainty in the retrieval and a higher likelihood of perceptual aliasing. In contrast, lower RS values suggest higher confidence, as a clear margin separates the best match from competing candidates. RS aggregates information from all top-k candidates rather than relying solely on the top two, making it more robust to outliers and more effective at capturing cases where multiple ambiguous matches exist.

B. Similarity Distribution (SD)

It measures the overall dispersion of similarity scores across the top-k candidates by computing the ratio of the median score to the highest score. This metric captures the global shape of the score distribution: lower SD values indicate large separations between the top match and remaining candidates, suggesting high confidence in the retrieval, while values approaching 1.0 reflect uniform score distributions where multiple candidates are equally competitive, indicating matching ambiguity. This is demonstrated in Fig. 2. Unlike methods that evaluate query-reference pairs in isolation, SD analyzes the collective distribution of top candidates, enabling more effective detection of perceptual aliasing scenarios where several visually similar but potentially incorrect matches compete with the true positive.

$$\text{SD} = \frac{s_{\text{median}}}{s_{\text{best}}} \tag{2}$$

The median is preferred over the mean as it is robust to the high-scoring false positives that skew VPR similarity distributions. A Mahalanobis-based alternative maybe considered but requires covariance estimation over only $k \approx 10$ samples, making it statistically unreliable without training data.

C. Statistical Uncertainty (SU)

It combines two complementary distributional properties of the similarity score distribution. These metrics are combined to form the final uncertainty estimate:

$$\text{SU} = \alpha \times \text{RS} + (1-\alpha) \times \text{SD} \tag{3}$$

where α controls the relative contribution of each component. This formula captures both local competitive dynamics among top candidates and global distributional characteristics of the similarity space. SU is not designed to maximize mean AUC-PR; rather, it serves as a validation-free safe default, eliminating the risk of selecting the worse individual metric when deployment context is unknown. Low SU scores typically correspond to confident matches with clear separation between candidates, while high SU scores indicate greater matching ambiguity, as illustrated in Fig. 2.

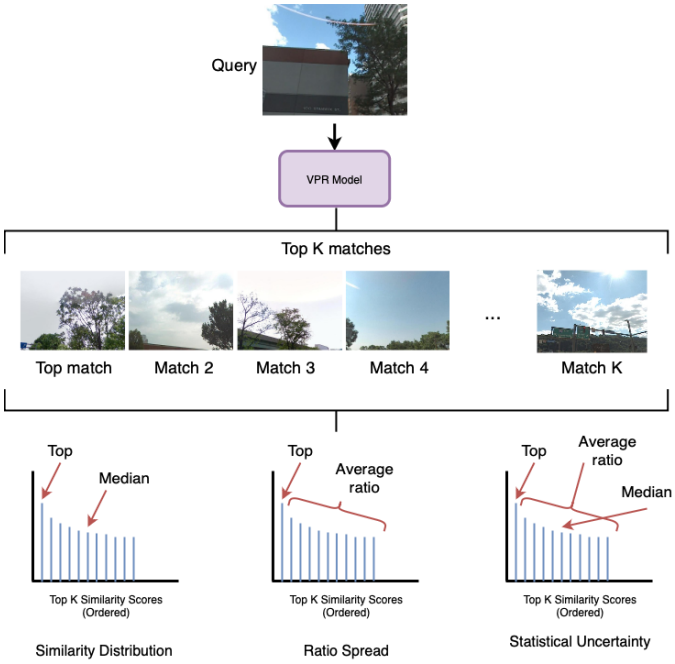

Fig. 2: Illustration of uncertainty metrics operating on similarity score distributions. RS captures competitive ambiguity through score clustering among top-k candidates, while SD measures match distinctiveness using the distance between maximum and median scores. SU combines both metrics to provide unified uncertainty estimation.

IV. Experimental Setup

This section provides details about the experimental setup.

A. Datasets and Methods

We evaluate nine VPR methods: Eigenplaces [29], Cosplace [30], ConvAP [31], MixVPR [15], AP GeM [32], SALAD [33], AnyLoc [34], CricaVPR [35], and NetVLAD [14] across six benchmark datasets, detailed in Table I.

B. Evaluation Metrics and Baselines

To ensure fair and interpretable evaluation, we assess uncertainty estimators independently from the retrieval backbone, isolating their ability to rank predictions by confidence. We evaluate uncertainty metrics through two complementary paradigms. Predictor Quality measures the intrinsic ability to discriminate between correct and incorrect matches using AUC-PR, where predictions within $d_{\max}$,the maximum spatial distance threshold used to determine a correct place match, of ground truth are labeled positive and confidence $C = -U(q, p_1)$, where U is uncertainty, enables binary classification, with q denoting the query image and p_1 the top-ranked retrieved reference image. End-to-End VPR Impact assesses practical deployment benefits by applying rejection threshold τ (accepting only if $U(q, p_1) < \tau$) and analyzing whether filtering uncertain predictions improves the precision-recall trade-off of the complete system.

We compare our approach against representative methods: (1) *Retrieval-based*: L2 Distance and Perceptual Aliasing (PA). (2) *Spatial*: SUE with $N = 10$ nearest neighbors, weighting slope $\sigma = 350$, and a maximum spatial distance cap of 500 m, as specified in the original SUE paper [20]. (3)

Dataset	Characteristics	Primary Challenge for Uncertainty Estimation
Pittsburgh 250k [36]	Large-scale urban environment with repetitive architectural structures and similar street layouts	Severe perceptual aliasing leading to high-confidence false matches
St. Lucia [37]	Suburban routes captured at different times with moderate viewpoint and appearance changes	Distinguishing weakly distinctive correct matches from genuine multi-way ambiguity
Nordland [38]	Fixed railway route observed under extreme seasonal variation	Identifying correct matches with low similarity due to appearance change rather than geometric confusion
Amstertime [39]	Urban environment with long-term temporal domain shift, including construction and illumination changes	Robust uncertainty estimation under distribution mismatch
Eynsham [40]	Highly repetitive residential scenes with visually similar locations	Extreme perceptual aliasing with multiple visually indistinguishable candidates
Tokyo24/7 [41]	Urban scenes captured across day-night cycles and varying weather conditions	Robustness to severe appearance change without conflating it with multi-way ambiguity

TABLE I: Dataset characteristics and associated challenges for uncertainty estimation in visual place recognition (VPR).

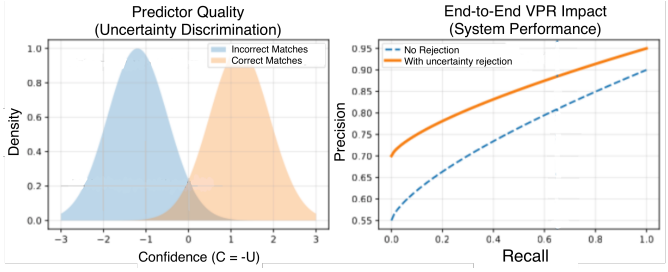

Fig. 3: Evaluation paradigms. (Left) Predictor Quality: Binary separation between correct and incorrect matches based on uncertainty scores. (Right) End-to-End Impact: VPR system Precision-Recall curve shifts upward as uncertainty rejection threshold tightens.

Learning-based: Bayesian Triplet Loss (BTL) with projection heads (linear layer for mean, Sequential block with Softplus for variance) producing distribution-based descriptors; triplet uncertainty computed by summing query and candidate variances with 0.5 margin. (4) *Geometric Verification (GV)*: Three pipelines with $U_{GV} = -c_{inliers}$, where U denotes the uncertainty estimate computed as the negative number of geometric inliers $c_{inliers}$ obtained after RANSAC verification. SIFT (FLANN matcher, Lowe's ratio 0.7, 10-match minimum, RANSAC threshold 5.0); DELF (seven scales 0.25–2.0, max 1,000 features, Brute-Force L2 matching, RANSAC homography); SuperPoint (SuperGlue outdoor weights, keypoint threshold 0.005, match threshold 0.2, MAGSAC with 5.0 reprojection threshold).

Our implementation of SUE uses a pretrained GeM ResNet-50 backbone rather than STUN-trained models from the original SUE paper [20]. While this introduces domain shift, yielding higher uncertainty when top-10 neighbors are visually similar but geographically scattered, it prioritizes computational efficiency and accessibility over the resource-intensive multi-stage teacher-student training. This design evaluates SUE's spatial uncertainty logic across different feature spaces while using existing high-quality extractors.

C. Configuration

Our metrics analyze similarity scores from the top 10 ($k = 10$) retrieved candidates using cosine distance via Faiss [42]. The SU score combines RS and SD with equal weighting ($\alpha = 0.5$). Experiments run on NVIDIA A100 GPUs with batch size 1 to simulate real-time robotic deployment. Results are averaged across three independent runs using standard dataset splits.

V. Results and Discussion

This section evaluates our uncertainty metrics in terms of performance, efficiency, and generalization across VPR systems.

A. Overall Performance Analysis

Table II evaluates predictor quality: how well each uncertainty metric distinguishes correct from incorrect matches, as measured by AUC-PR. A score closer to 1.0 indicates a more effective predictor, this measures the quality of the confidence estimates themselves.

Our three metrics consistently outperform existing retrieval-based approaches (L2, PA), learning-based methods (BTL) and spatial methods (SUE), achieving higher AUC-PR scores across nearly all VPR architectures. In particular, SU achieves the best overall results with a mean AUC-PR of 0.92, demonstrating robust uncertainty estimation across diverse visual environments and architectures.

The Precision-Recall curves across six configurations are shown in Fig. 4. Our uncertainty metrics consistently outperform traditional approaches, achieving substantially higher AUC than L2 distance and competitive performance with geometric verification methods. Performance gains are especially notable in challenging conditions: on Tokyo24/7's day-night variations and Nordland's seasonal changes, our methods maintain high precision at elevated recall while baselines degrade. Strong performance across CricaVPR and SALAD demonstrates architecture-agnostic applicability.

B. Cross-Method Generalization

Fig. 5 visualizes the performance gains (in AUC-PR) of our metrics relative to baseline uncertainty methods across different VPR architectures and datasets. Our metrics consistently show positive gains, particularly on challenging datasets like Amstertime (viewpoint changes) and Nordland (seasonal variation), demonstrating robustness from CNN-based (NetVLAD) to Transformer-based models (Eigenplaces, SALAD). Compared to existing approaches, our methods address key limitations: geometric verification techniques (SuperPoint) achieve accuracy but are computationally prohibitive

Uncertainty Metric	Eigenplaces	Cosplace	ConvAP	MixVPR	APGeM	SALAD	AnyLoc	Crica	NetVLAD	Mean	Time
Baseline Model (BM)	0.91	0.9	0.87	0.91	0.62	0.94	0.81	0.87	0.71	0.84	–
BM + (DUE) BTL	0.24	0.25	0.17	0.17	0.13	0.33	0.26	0.13	0.10	0.20	0.52
BM + (RUE) L2	0.89	0.90	0.85	0.90	0.62	0.94	0.82	0.87	0.71	0.83	0.01
BM + (RUE) PA Score	0.93	0.93	0.89	0.92	0.74	**0.97**	0.88	0.93	0.78	0.88	0.01
BM + (RUE) RS	0.94	**0.95**	**0.92**	**0.94**	0.80	0.96	0.90	0.94	0.82	0.91	0.01
BM + (RUE) SD	**0.95**	0.94	**0.92**	**0.94**	**0.82**	0.96	0.90	0.93	**0.83**	0.91	0.02
BM + (RUE) SU	**0.95**	**0.95**	**0.92**	**0.94**	**0.82**	**0.97**	**0.91**	**0.96**	**0.83**	**0.92**	0.03
BM + SUE	0.68	0.68	0.71	0.81	0.48	0.88	0.73	0.88	0.53	0.71	0.05
BM + (GV) SIFT	0.72	0.72	0.67	0.66	0.38	0.83	0.66	0.72	0.42	0.64	139.86
BM + (GV) DELF	0.79	0.78	0.75	0.85	0.72	0.93	0.81	0.87	0.74	0.80	78.54
BM + (GV) SuperPoint	**0.95**	**0.95**	**0.98**	**0.95**	**0.93**	0.96	**0.94**	0.95	**0.94**	**0.95**	84.11

TABLE II: Comparative evaluation of uncertainty metrics across VPR methods. BM reports baseline Recall@1 *without* uncertainty rejection, while subsequent rows report AUC-PR scores measuring discrimination between correct and incorrect matches. Note that BM (Recall@1) and BM+L2 (AUC-PR) measure different quantities. Higher AUC-PR indicates better uncertainty estimation (bold = best). The final column shows processing time per query in seconds.

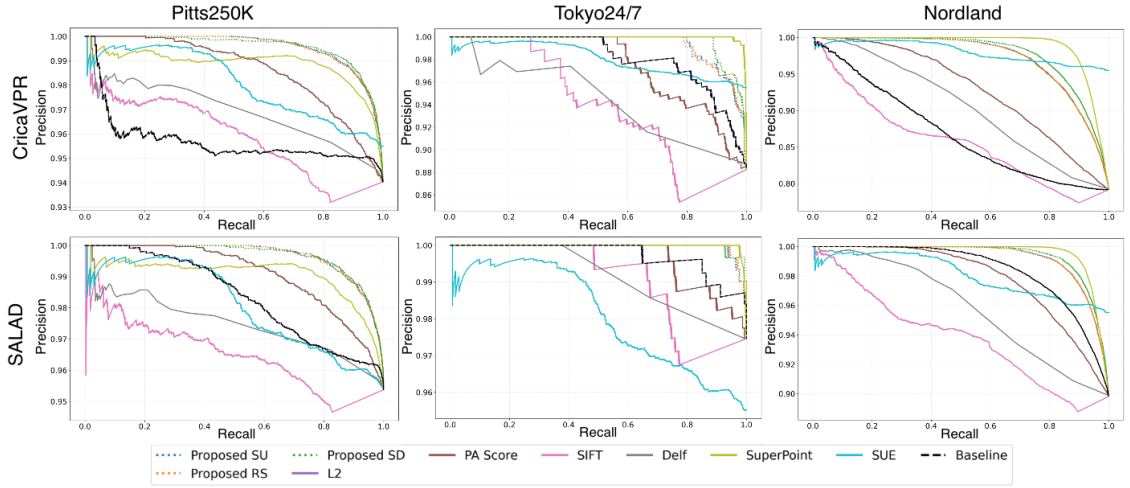

Fig. 4: Precision-Recall curve comparison of uncertainty estimation methods across diverse VPR scenarios. We evaluate two state-of-the-art VPR methods (CricaVPR and SALAD) on three challenging benchmarks: Pitts250K, Tokyo24/7, and Nordland. Each plot compares our proposed uncertainty metrics (Proposed SU, RS, and PA) against the baseline (no uncertainty filtering) and established methods including L2 distance, PA Score, SU, Delf, Sift, SuperPoint, and SUE. BTL is omitted due to significantly poor performance that obscures meaningful comparison.

for real-time use, retrieval-based baselines (L2, PA) are efficient but architecturally inconsistent, and learning-based methods (BTL, SUE) either underperform or require architecture-specific retraining. Our metrics combine accuracy and real-time efficiency for practical VPR uncertainty estimation.

C. Per Challenge Stratification

Table III reports AUC-PR stratified by illumination condition on Tokyo24/7 and seasonal domain shift on Nordland. All three metrics remain strong across conditions, with the most pronounced degradation under Fall→Winter (AUC-PR≈ 0.91, Recall@1 to 62%), confirming that score-distribution uncertainty remains informative precisely where the VPR backbone struggles most.

Setting	Condition	Recall@1	RS	SD	SU
Day/Night	Day	0.95	0.99	0.99	0.99
	Night	0.83	0.99	0.98	0.99
	Overall	0.89	0.99	0.99	0.99
Seasonal	Fall→Spr	0.90	0.99	0.99	0.99
	Fall→Sum	0.97	0.99	0.99	0.99
	Fall→Win	0.62	0.91	0.91	0.91

TABLE III: Stratified AUC-PR by illumination condition (Tokyo24/7) and seasonal domain shift (Nordland), reference database fixed to Fall.

D. Failure Mode Analysis

We partition all test queries into four quadrants based on whether RS and SD exceed the per-dataset median and measure the failure rate in each. As shown in Fig. 6, confident retrievals (Low RS / Low SD) are nearly error-free while uniformly ambiguous queries (High RS / High SD) show the highest failure rates. The off-diagonal quadrants behave differently, perceptual aliasing cases (High RS / Low SD) consistently fail more than spatially coherent cases (Low RS / High SD), confirming that RS and SD flag distinct failure modes rather than redundant information.

E. Computational Efficiency

The proposed metrics are highly efficient, requiring less than 0.03 ms per query (Table II). This is orders of magnitude faster than geometric verification methods such as SuperPoint (84 ms), making them suitable for real-time robotic applications with limited computational resources. The efficiency comes from performing lightweight statistical analysis directly on pre-computed similarity scores, avoiding feature extraction or matching overhead.

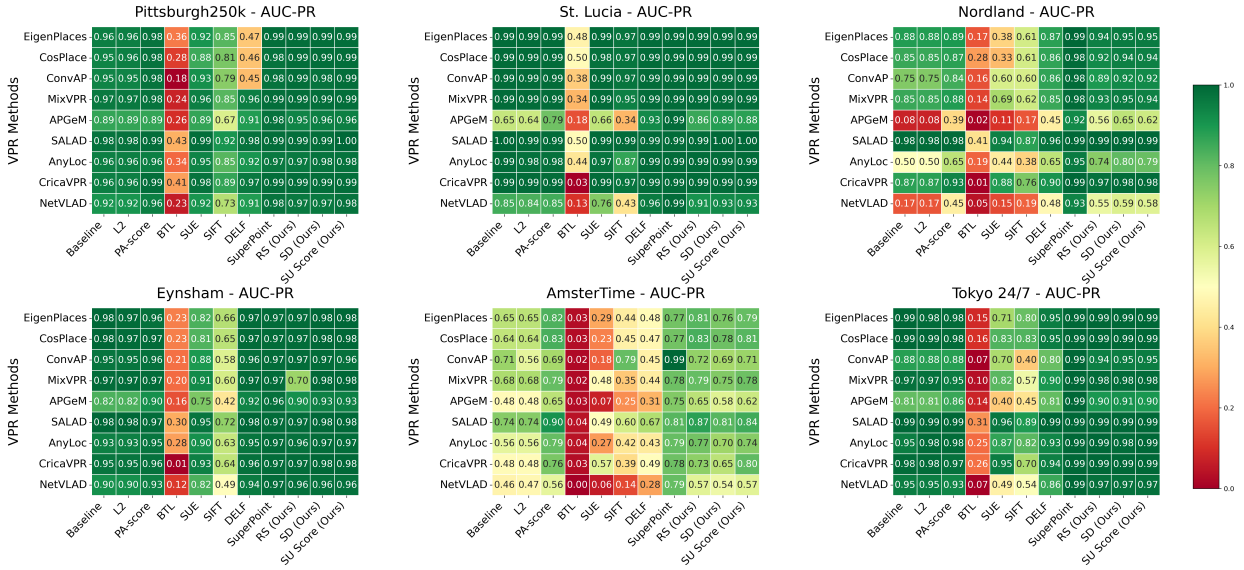

Fig. 5: AUC-PR comparison of uncertainty estimation methods across diverse VPR scenarios. Each plot compares our proposed uncertainty metrics against the baseline and established methods. AUC values are reported in the box for each method.

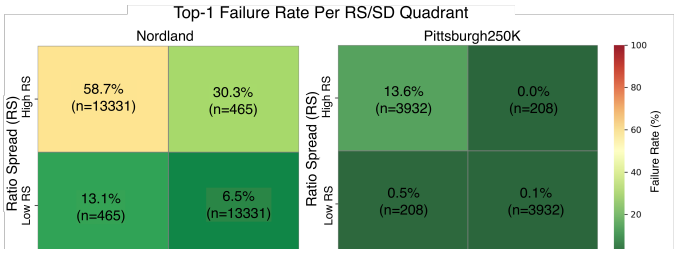

Fig. 6: 2×2 query partitioning by RS and SD (thresholded at per-dataset median). Asymmetric off-diagonal failure rates confirm RS and SD capture distinct rather than redundant failure modes.

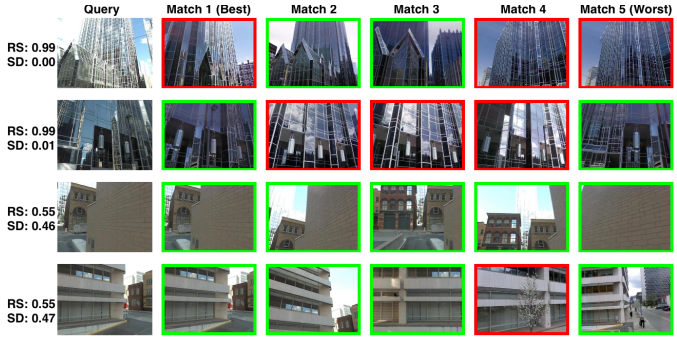

Fig. 7: Complementarity of RS and SD. Green borders indicate correct matches, red borders failures. Rows 1–2: High RS / Low SD highlights perceptual aliasing of visually similar but geographically scattered facades. Rows 3-4: High SD / Low RS shows spatially close but visually diverse retrievals. Off-diagonal failure patterns, Fig. 6 confirm complementary modes, motivating their combination in SU.

VI. Ablation Study

This ablation studies the complementary nature of RS and SD, the effect of α weighting in the SU equation, and robustness of the top- k value.

A. RS vs SD Complementarity Analysis

RS and SD operate on top-k score ratios with similar performance (Fig. 5) but capture different failure modes. Their strong positive correlation ($\rho = 0.98$, $p < 0.001$) reflects that both metrics respond to retrieval difficulty, yet their failure behaviors diverge in the off-diagonal quadrants, as quantified in Fig. 6. Fig. 7 shows distinct failure patterns: high RS with low SD identifies perceptual aliasing, visually identical but geographically scattered candidates, while high SD with low RS captures spatially close but visually diverse retrievals where appearance varies across viewpoints or conditions. Their combination in SU jointly addresses failures neither metric captures alone.

B. SU Alpha Weighting Study

SU performance for $\alpha \in [0.0, 1.0]$ across APGeM and EigenPlaces, shown in Fig. 8. While most datasets exhibit stable performance, challenging cases like Amstertime and Nordland prefer either SD or RS depending on method. We select $\alpha = 0.5$ because: (1) RS and SD are strongly correlated overall yet diverge in their sensitivity to specific failure modes, as evidenced by the contrasting off-diagonal failure rates in Fig. 6, justifying equal weighting; (2) $\alpha = 0.5$ performs consistently well without validation tuning, avoiding drops when one metric dominates. Intuitively, their equal-weight combination provides a smoother uncertainty surface: when one metric fails to flag an ambiguous case, the other partially compensates, reducing the systematic blind spots that either metric exhibits alone.

Table IV provides direct evidence of SU's worst-case performance guarantee. Specifically, SU ensures performance no worse than the poorer of its two underlying metrics (RS and SD), effectively providing an insurance mechanism against metric misselection. Across all VPR methods, SU matches or exceeds the worse of RS and SD, confirming that it eliminates the risk of selecting an underperforming metric while maintaining near-peak performance.

As RS and SD diverge in sensitivity to specific failure

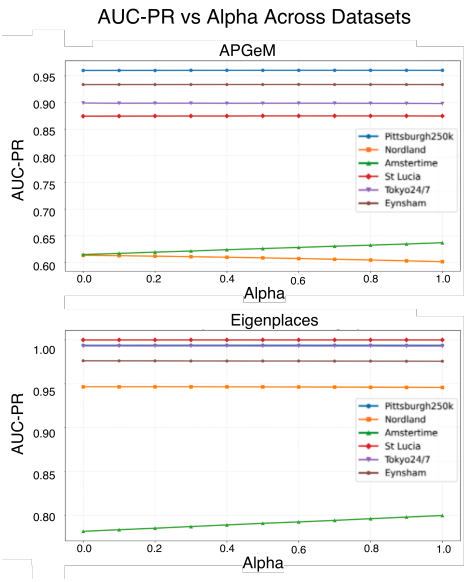

Fig. 8: AUC-PR performance as a function of α weighting parameter (Equation 3) across multiple datasets. (Left) Results using APGeM as the VPR method. (Right) Results using EigenPlaces as the VPR method. Each line represents a different dataset, showing how the combined uncertainty metric SU performs across varying contributions of RS and SD components.

Method	Best(RS,SD)	Worst(RS,SD)	SU	ΔWorst
Eigenplaces	0.95	0.94	0.95	+0.01
APGeM	0.82	0.80	0.82	+0.02
NetVLAD	0.83	0.82	0.83	+0.01
AnyLoc	0.90	0.90	0.91	+0.01
Mean	0.91	0.91	0.92	+0.01

TABLE IV: SU worst-case insurance across VPR methods. SU consistently matches or exceeds the worst individual metric at no cost to the best, confirming its role as a validation-free default.

modes (Fig. 6), their equal-weight combination compensates for each other's blind spots, with $\alpha = 0.5$ being the natural choice to avoid reintroducing dependence on validation data.

C. K-Value Robustness Analysis

This ablation study examines the influence of the top-k candidates used in our uncertainty metrics, evaluated on the Pittsburgh250k dataset across multiple VPR methods, with findings that generalize consistently to other datasets.

We analyzed the influence of the number of top candidates (k) on Pittsburgh250k and Eynsham. As shown in Fig. 9, both SD and RS stabilize around $k = 10$, plateauing beyond that. This occurs because once k captures the perceptual aliasing neighborhood, additional candidates contribute diminishing information to the median and ratio statistics, a pattern confirmed on the highly repetitive Eynsham dataset. We note that in high-frame-rate or densely sampled databases, a small k may insufficiently cover the aliasing neighborhood; practitioners should validate k on a small held-out sample in such settings.

VII. Conclusions and Future Work

This paper introduced three training-free uncertainty metrics for VPR: Ratio Spread (RS), Similarity Distribution (SD), and their unified combination, Statistical Uncertainty (SU). Our key contribution is a method-agnostic approach that requires no training, architectural modifications, or validation data for metric selection, while adding negligible computational overhead (<0.03 ms per query).

Despite strong overall performance, the proposed metrics have identifiable failure boundaries. In environments with globally repetitive structures, such as long tunnels or symmetric architectural facades, the top-k score distribution offers little discriminative signal because all candidates are visually near-identical; geometric verification remains necessary in such extreme cases. Score-distribution methods also degrade under database sparsity, where the correct match may simply be absent, causing the metrics to flag correct retrievals as uncertain. Finally, severe sensor degradation can cause uniform score collapse that masks meaningful distribution structure. We believe these boundaries motivate future hybrid approaches combining lightweight distributional uncertainty with selective geometric verification triggered only when distribution-based confidence falls below a threshold.

References

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] 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, 2016.

[3] S. Schubert, P. Neubert, S. Garg, M. Milford, and T. Fischer, "Visual place recognition: A tutorial [tutorial]," *IEEE Robotics & Automation Magazine*, vol. 31, no. 3, p. 139–153, Sep. 2024. [Online]. Available: http://dx.doi.org/10.1109/MRA.2023.3310859

[4] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?" in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, ser. IJCAI-2021. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, p. 4416–4425. [Online]. Available: http://dx.doi.org/10.24963/ijcai.2021/603

[5] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, "A survey on Visual-Based Localization: On the benefit of heterogeneous data," *Pattern Recognition*, vol. 74, pp. 90 – 109, Feb. 2018. [Online]. Available: https://hal.science/hal-01744680

[6] S. Hausler, T. Fischer, and M. Milford, "Unsupervised complementary-aware multi-process fusion for visual place recognition," 2021. [Online]. Available: https://arxiv.org/abs/2112.04701

[7] F. Warburg, M. Jørgensen, J. Civera, and S. Hauberg, "Bayesian triplet loss: Uncertainty quantification in image retrieval," 2021. [Online]. Available: https://arxiv.org/abs/2011.12663

[8] K. Cai, C. X. Lu, and X. Huang, "Stun: Self-teaching uncertainty estimation for place recognition," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 6614–6621.

[9] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–, 11 2004.

[10] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," 2018. [Online]. Available: https://arxiv.org/abs/1612.06321

[11] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 337–33 712.

[12] P. Yin, J. Jiao, S. Zhao, L. Xu, G. Huang, H. Choset, S. Scherer, and J. Han, "General place recognition survey: Towards real-world autonomy," 2025. [Online]. Available: https://arxiv.org/abs/2405.04812

[13] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int. J. Rob. Res.*, vol. 27, no. 6, pp. 647–665, Jun. 2008.

[14] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," 2016. [Online]. Available: https://arxiv.org/abs/1511.07247

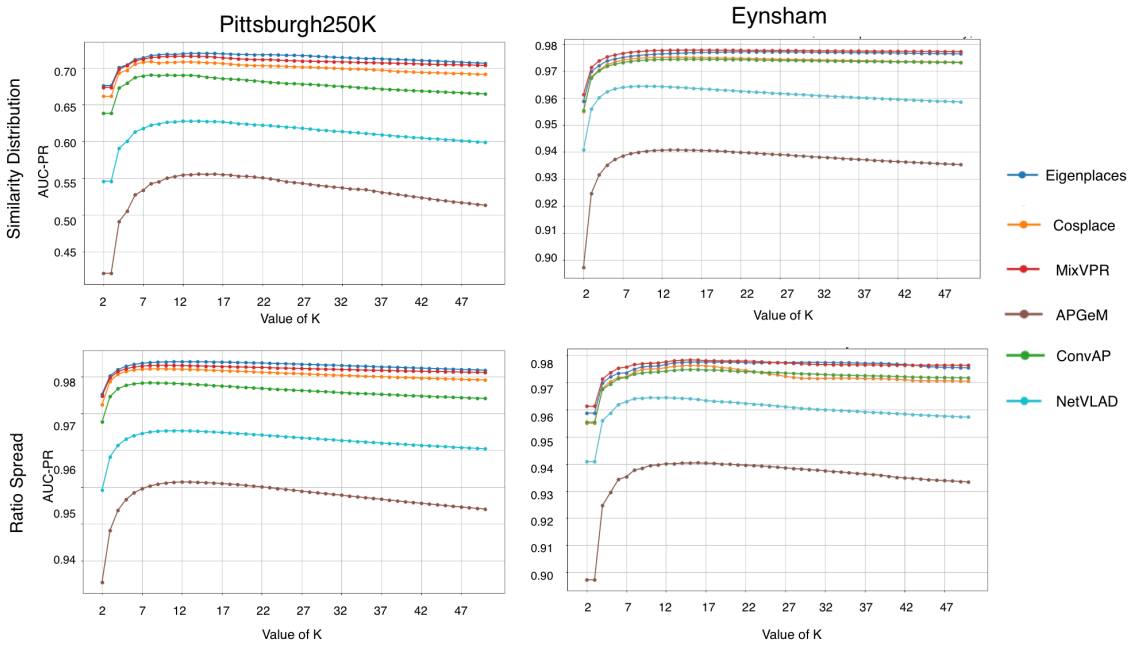

Fig. 9: Comparison of SD and RS uncertainty metrics across k values for VPR methods on Pittsburgh250k (Top) and Eynsham (Bottom). AUC-PR results show both SD and RS provide more consistent performance across k values.

[15] A. Ali-bey, B. Chaib-draa, and P. Giguère, "Mixvpr: Feature mixing for visual place recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 2998–3007.

[16] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," 2022. [Online]. Available: https://arxiv.org/abs/2201.02001

[17] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, "Generalized contrastive optimization of siamese networks for place recognition," 2023. [Online]. Available: https://arxiv.org/abs/2103.06638

[18] J. Revaud, J. Almazan, R. Rezende, and C. D. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019.

[19] A. Ali-bey, B. Chaib-draa, and P. Giguère, "Gsv-cities: Toward appropriate supervised visual place recognition," *Neurocomputing*, vol. 513, p. 194–203, Nov. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2022.09.127

[20] M. Zaffar, L. Nan, and J. F. P. Kooij, "On the estimation of image-matching uncertainty in visual place recognition," 2024. [Online]. Available: https://arxiv.org/abs/2404.00546

[21] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5580–5590.

[22] M. Hasan, A. Khosravi, I. Hossain, A. Rahman, and S. Nahavandi, "Controlled dropout for uncertainty estimation," 2022. [Online]. Available: https://arxiv.org/abs/2205.03109

[23] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," 2017. [Online]. Available: https://arxiv.org/abs/1612.01474

[24] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," 2018. [Online]. Available: https://arxiv.org/abs/1802.10501

[25] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," 2018. [Online]. Available: https://arxiv.org/abs/1610.02136

[26] F. Neha, D. Bhati, D. K. Shukla, and M. Amiruzzaman, "From classical techniques to convolution-based models: A review of object detection algorithms," 2024. [Online]. Available: https://arxiv.org/abs/2412.05252

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014. [Online]. Available: https://arxiv.org/abs/1311.2524

[28] J. Lee, T. Miyanishi, S. Kurita, K. Sakamoto, D. Azuma, Y. Matsuo, and N. Inoue, "Citynav: A large-scale dataset for real-world aerial navigation," 2025. [Online]. Available: https://arxiv.org/abs/2406.14240

[29] 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 (ICCV)*, October 2023, pp. 11 080–11 090.

[30] 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 (CVPR)*, June 2022, pp. 4878–4888.

[31] A. Ali-bey, B. Chaib-draa, and P. Giguère, "Gsv-cities: Toward appropriate supervised visual place recognition," *Neurocomput.*, vol. 513, no. C, p. 194–203, Nov. 2022. [Online]. Available: https://doi.org/10.1016/j.neucom.2022.09.127

[32] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, 2019.

[33] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," 2024. [Online]. Available: https://arxiv.org/abs/2311.15937

[34] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," 2023. [Online]. Available: https://arxiv.org/abs/2308.00688

[35] F. Lu, X. Lan, L. Zhang, D. Jiang, Y. Wang, and C. Yuan, "Cricavpr: Cross-image correlation-aware representation learning for visual place recognition," 2024. [Online]. Available: https://arxiv.org/abs/2402.19231

[36] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla, "Visual place recognition with repetitive structures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[37] M. J. Milford and G. F. Wyeth, "Mapping a suburb with a single camera using a biologically inspired slam system," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1038–1053, 2008.

[38] S. Skrede, "Nordland dataset," 2013, accessed: March 12, 2026. [Online]. Available: https://bit.ly/2QVBOym

[39] B. Yildiz, S. Khademi, R. M. Siebes, and J. van Gemert, "Amstertime: A visual place recognition benchmark dataset for severe domain shift," 2022. [Online]. Available: https://arxiv.org/abs/2203.16291

[40] M. Cummins and P. Newman, "Highly scalable appearance-only slam - fab-map 2.0," 06 2009.

[41] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 257–271, 2018.

[42] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," 2024.