Journal of Applied Computational Intelligence

eISSN: 3136-6937

Vol. 4, No. 2, April, 2026

Season Invariant NetVLAD Loop Closure Detection through Semantic Hard Negative Mining

Kamil Wysocki[1,*], Adrian Ostrowski[1]

[1] *Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, 44-100, Poland*

Received 29 January 2026; Revised 31 May 2026; Accepted 03 June 2026; Published 09 June 2026

Abstract

Seasonal appearance changes complicate visual loop closure because corresponding locations can become less similar than geographically unrelated scenes with comparable semantic layouts. This paper develops a semantic hard negative mining framework for a season invariant NetVLAD descriptor. The framework combines confidence weighted structural pooling, geographically guarded candidate selection, and a gradual transition from moderately difficult to difficult training pairs. Semantic similarity identifies informative distractors, while a spatial exclusion region and an uncertainty buffer reduce the risk of treating revisited locations as negatives. The evaluation examines retrieval accuracy, component contributions, and computational resource requirements under a controlled protocol. Across four seasonal transition scenarios, the proposed configuration attains top ranked retrieval recalls of 88.4%, 86.1%, 83.7%, and 81.8%, compared with 79.1%, 76.8%, 73.5%, and 70.2% for the matched NetVLAD baseline. Mean recall therefore increases from 74.9% to 85.0%. The component study assigns 78.0% to structural pooling alone and 81.2% to guarded semantic mining, exposing the complementary roles of feature weighting and sample selection. The configured retrieval pipeline uses a 4096-dimensional descriptor and a median latency of 35.0 ms. Threshold and perturbation analyses examine precision sensitivity, semantic uncertainty, storage growth, and difficult failure cases. The evaluation relates structural feature weighting and training sample selection to retrieval accuracy, verification behavior, and deployment requirements. The mean top-ranked retrieval gain is 10.1 percentage points; this descriptor-level comparison is assessed separately from geometric acceptance of loop candidates.

Keywords: Seasonal Place Recognition, Semantic Hard Negative Mining, NetVLAD Descriptor, Loop Closure Detection, Structural Feature Pooling

1. Introduction

Visual place recognition supports mobile robot localization by identifying previously observed places from image evidence [1]. Recent methods address severe appearance variation by combining global retrieval with local feature matching in a unified recognition pipeline [2]. Universal place recognition approaches combine self-supervised visual representations with feature aggregation to improve transfer across diverse environments [3]. Research on localization integrity uses query rejection and trajectory-based verification to improve the reliability of place recognition for robot navigation [4]. This distinction is particularly relevant in repeated street layouts, where a descriptor may retrieve a visually plausible but geographically incorrect building frontage. A missed revisit allows accumulated drift to persist, whereas an incorrect revisit can introduce an inconsistent constraint into the map. Vegetation growth, illumination variation, and temporary occlusion can affect several sources of image evidence simultaneously, making seasonal recognition a problem of interacting appearance changes. A recognition system intended for repeated deployment must account for these interactions across the conditions encountered during operation. The relevant engineering objective

* Corresponding author: wysocki.ka@stud.polsl.pl

http://dx.doi.org/10.68406/jaci.2026.vol4iss2no11:150-165

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

therefore includes stable retrieval performance, reliable rejection, and compatibility with downstream geometric verification. A useful descriptor must preserve persistent evidence, reject lookalike places, and expose uncertainty at the interface with the mapping system. In seasonal operation, the same physical location can retain its road geometry and building layout while foliage, surface cover, and illumination alter the evidence available to the descriptor.

Aggregated visual descriptors provide a compact interface between image interpretation and database retrieval. Multi-scale NetVLAD aggregation incorporates intermediate backbone features alongside bottleneck features to improve place descriptors [5]. In this setting, NetVLAD supplies a differentiable pooling operation that can be trained with place related supervision. However, a compact global representation can retain correlations that are predictive within one traversal and unreliable across seasons. Cross-environment feature enhancement has used generated condition variations to strengthen representation learning against seasonal and illumination changes [6]. A model can assign disproportionate importance to vegetation texture, road brightness, or sky color when those signals simplify the training objective. Semantic information offers a possible source of persistence because broad scene categories may remain recognizable after local textures change. Semantic boundary supervision directs feature extraction toward structural regions that remain informative under adverse appearance conditions [7]. Semantic similarity describes scene composition, while place identity requires location specific evidence. Different residential streets can contain almost identical proportions of buildings, road, and vegetation. Appearance invariance must retain sufficient spatial discrimination to distinguish geographically unrelated locations. The desired invariance must preserve the differences needed to decide whether two observations originate from the same part of the route. The cited recent variants illustrate developments in aggregation and robustness; the present comparison concerns their design principles rather than a historical priority claim about the original pooling layer.

Training pair selection has a direct influence on which distinctions a descriptor learns. Guan et al. [8] introduced a hardness-aware metric loss that weights negative examples according to their similarity to the query and the stage of training. Under seasonal change, a useful difficult negative combines a similar semantic arrangement with independently established geographic separation. Weak positional supervision can misclassify a genuine revisit as a negative pair. Sample-wise weighted distillation prioritizes reliable place image pairs while suppressing the influence of noisy pairs [9]. Spatial uncertainty and adjacent viewpoints make this issue acute on routes with parallel lanes, narrow side roads, and imprecise positioning. Adaptive mining has been integrated with continual place recognition to balance performance within an environment and generalization across environments [10]. This provides a rationale for delaying the strongest semantic distractors until the representation can separate basic place identity. Existing lines of work leave an engineering gap between semantic persistence, negative validity, and training schedule design. A seasonal loop closure model needs a sampling policy that exploits semantic resemblance while retaining an explicit means of rejecting ambiguous labels. Evaluation must also distinguish descriptor retrieval from the subsequent geometric verification of loop candidates. Such a policy must maintain candidate diversity across independent route segments.

The interaction between structural feature weighting and semantic sample selection calls for an evaluation that connects representation behavior with the practical requirements of seasonal loop closure. Controlled comparisons can examine this interaction through matched backbones, descriptor sizes, training exposure, and candidate verification settings. Spatial clustering has also been used to strengthen the consistency of NetVLAD-based descriptors under aerial viewpoint variation [11]. Deployment-oriented place recognition studies assess model and map memory together with feature extraction and retrieval latency [12]. These considerations motivate the present design, which couples structural feature weighting with geographically guarded semantic hard negative mining and a gradual difficulty schedule. The paper develops a reproducible evaluation protocol for examining the interaction between descriptor aggregation, negative selection, and loop acceptance. The remaining sections develop the method and its evaluation in sequence. Section 2 reviews representation aggregation, semantic supervision, and sample selection, identifying the unresolved interface between them. Section 3 specifies the model, its mining rules, the training objective, and its integration with candidate verification. Section 4 defines a reproducible evaluation protocol and examines retrieval performance, component effects, operating thresholds, computational behavior, and failure conditions. Section 5 states the methodological conclusions, limitations, and directions for validation on independently collected seasonal traversals. Seasonal invariance denotes the objective of preserving place discrimination across seasonal appearance changes within the stated evaluation conditions. Transfer to unseen routes, sensors, and weather regimes requires separate validation. The study develops guarded semantic candidate selection, links it to structurally weighted aggregation, and specifies component and acceptance analyses for examining their interaction. These contributions concern training supervision, descriptor construction, and evaluation design, respectively.

2. Related Work

2.1. Global Descriptors under Seasonal Appearance Change

Global place descriptors compress image evidence into a vector suitable for nearest neighbor retrieval. Clustering-based weighted feature maps and optimized residual aggregation have been used to reduce redundant image information

in global place descriptors [13]. Their usefulness depends on whether local assignments remain sufficiently stable under changes in viewpoint and environmental condition. A seasonal transformation can modify many low-level measurements without changing the physical scene, so geographic correspondence requires additional place specific evidence. Learned aggregation addresses part of this problem by adapting local descriptors and pooling parameters to the retrieval objective. Decoder-based aggregation methods use learnable queries to integrate contextual visual features into a global place representation [14]. Its fixed length vector supports efficient database retrieval before local matching is applied to selected candidates. Nevertheless, training supervision determines which visual differences are suppressed and which remain discriminative. A descriptor trained on convenient within season correspondences may preserve season specific properties that become unreliable during later revisits.

Multi scale and region aware representations seek to retain more information about the spatial distribution of scene evidence. Multi-scale patch selection and geometric verification can refine candidates initially retrieved using global image descriptors [15]. Such designs can distinguish images that share broad scene categories while differing in the arrangement of architectural details. Their additional matching stage, however, creates a resource tradeoff when the database grows or the robot must respond within a limited processing interval. Domain-invariant representation methods emphasize static object information and jointly constrain multiple positive and negative observations during training [16]. Their evaluation depends on whether held out conditions also differ in route geometry, sensor configuration, or class distribution. These factors are relevant to seasonal recognition because a favorable split can influence the apparent gain attributed to invariance. The present problem requires a representation that remains compact while learning specifically from geographically distinct scenes that have similar semantic composition. The selection of those scenes as training negatives remains a separate sampling design requirement.

2.2. Semantic Persistence and Confidence Weighting

Semantic guidance introduces information about scene content beyond raw texture. Semantic and depth inpainting methods suppress dynamic objects while retaining scene geometry for place recognition [17]. This distinction also applies to seasonal change, where category reliability varies with environmental conditions. Vegetation is physically static at the level of a tree trunk but highly variable at the level of foliage appearance. Roads remain meaningful landmarks while their visible material can be concealed by snow. Class specific reliability weights account for these differences. Multimodal place recognition methods use attention-based fusion to exchange complementary information between image and point cloud features [18]. For seasonal recognition, uncertainty can reflect both an ambiguous pixel and a systematic domain shift in the segmentation model. A confidence score consequently serves as a continuous control on semantic influence, with its reliability established through calibration.

Semantic scene signatures also provide a compact measure of environmental resemblance. Bao et al. [19] combined semantic gradient features and scene structure with residual aggregation and local spatial verification for long-term place recognition. Such similarity identifies distractors whose location identity must be established through independent evidence. Similar arrangements of roads, buildings, and trees recur at geographically distinct locations along a route. An effective semantic mining policy should expose those repeated layouts as negative evidence only after position uncertainty has been considered. The research gap concerns the joint treatment of semantic resemblance and supervision validity. Structural weighting can reduce unstable evidence in the descriptor, while semantic resemblance can identify difficult training examples; both functions require an independent geographic guard to establish pair validity. Preserving this distinction also improves interpretability when an experiment reveals a loss of discrimination between neighboring locations. The proposed integration assigns separate roles to structural pooling, semantic distractor identification, geographic validity screening, and progressive difficulty control. Its contribution concerns their coordinated use within a compact loop-closure front end.

2.3. Hard Negatives and Reliable Loop Acceptance

Hard negative mining targets examples that violate or nearly violate the intended ordering of positive and negative distances. Ranking objectives can emphasize difficult candidates through losses designed to provide more informative gradients for hard samples [20]. This distinction matters when place labels rely on approximate positions with incomplete geometric correspondence. A nearest negative in descriptor space may be an unannotated positive, a nearby viewpoint, or a genuinely distinct location. Checking supervision quality preserves valid correspondences during difficult pair selection. Training-free uncertainty measures use retrieval score separation and competition among leading candidates to assess ambiguous place matches [21]. For route-based data, an explicit distance buffer offers a practical interpretation of that principle, provided the buffer is derived from positioning uncertainty and fixed using a separate development partition.

Training protocols determine how spatial and temporal observations contribute to representation learning. Joint image and sequence training uses a multitask learning protocol to learn compact representations from isolated images and temporal observations [22]. A seasonal descriptor may benefit from moderate negatives during initialization and

semantic lookalikes after basic geographic discrimination develops. The schedule itself is an experimental factor: more frequent mining changes computational cost and the distribution of training examples. At deployment, a descriptor remains a proposal mechanism even when its ranking improves. Local positional graphs encode the spatial context of image features to improve candidate matching within a hierarchical recognition pipeline [23]. Their operation must be evaluated separately from ranking, since a high scoring false candidate can be rejected while a correct candidate may fail verification under low texture. The unresolved requirement is a complete interface linking reliable mining, compact representation, and conservative acceptance. This motivates a controlled design in which each training component has a defined role and each online decision retains an explicit rejection path.

3. Semantic Mining and Descriptor Learning

3.1. Confidence Weighted Structural Aggregation

The proposed front end receives a color image resized to 640 × 384 pixels. A convolutional encoder produces a spatial grid of local descriptors, while a frozen semantic predictor assigns category probabilities and confidence values to the corresponding locations. The semantic outputs support spatial pooling and training signature construction, whereas geographic proximity and view overlap establish place correspondence. Structural category reliability is initialized with high values for building boundaries and permanent road furniture, intermediate values for road surfaces, and lower values for sky and moving objects. Vegetation retains a nonzero contribution because its spatial structure can remain informative on rural routes. These reliability settings are treated as development parameters whose transferability requires evaluation across environmental conditions. Backbone architecture, segmentation checkpoint, preprocessing normalization, and feature-grid resolution are implementation dependencies that must accompany a reproducible trained-model release. Their identities are not established by descriptor dimensionality alone.

At each spatial location, a bounded weight combines semantic class reliability with prediction confidence. The lower limit of 0.10 preserves visual evidence when semantic predictions are uncertain, while the upper limit of 1.00 controls the contribution of individual regions. Local descriptor normalization before weighting separates activation magnitude from semantic reliability. The aggregation equation incorporates these weights into the residual assigned to each visual cluster: Class reliability expresses the expected persistence of a category, while semantic confidence reflects the predictor's local output. The bounded pooling weight combines these distinct inputs; its calibration must be assessed separately from the clipping limits.

$$\mathbf{V}_k = \sum_{i=1}^{N} w_i a_{ik} (\mathbf{x}_i - \mathbf{c}_k) \tag{1}$$

Here, $\mathbf{x}_i$ denotes the local descriptor at spatial location i, $\mathbf{c}_k$ is the center of cluster k, and a_{ik} is the corresponding soft assignment. The assignments sum to one across clusters for each location, and w_i denotes the bounded structural weight. The equation allows reliable structural regions to contribute more strongly while retaining a residual contribution from less certain content. Each aggregated cluster vector undergoes intra-normalization before concatenation. A learned projection reduces the concatenated representation to 4096 components, followed by unit normalization for database retrieval. A small numerical stabilizer handles clusters with negligible aggregate support. The semantic predictor remains frozen throughout descriptor training, and category reliability settings selected on the development partition are retained across seasonal transitions.

Figure 1 presents the relationship between the descriptor path, semantic weighting, and the training signature branch. Frozen and trainable operations are distinguished so that the computational boundary remains explicit. Semantic prediction and spatial weighting operate during inference and therefore enter the online extraction budget. Semantic signatures and the negative queue support offline training. This allocation permits a resource comparison with the unweighted baseline that accounts for the additional semantic inference cost. Online cost includes semantic prediction, spatial weighting, visual encoding, aggregation, retrieval, and verification. Signature queue construction and negative mining belong to training resource accounting.

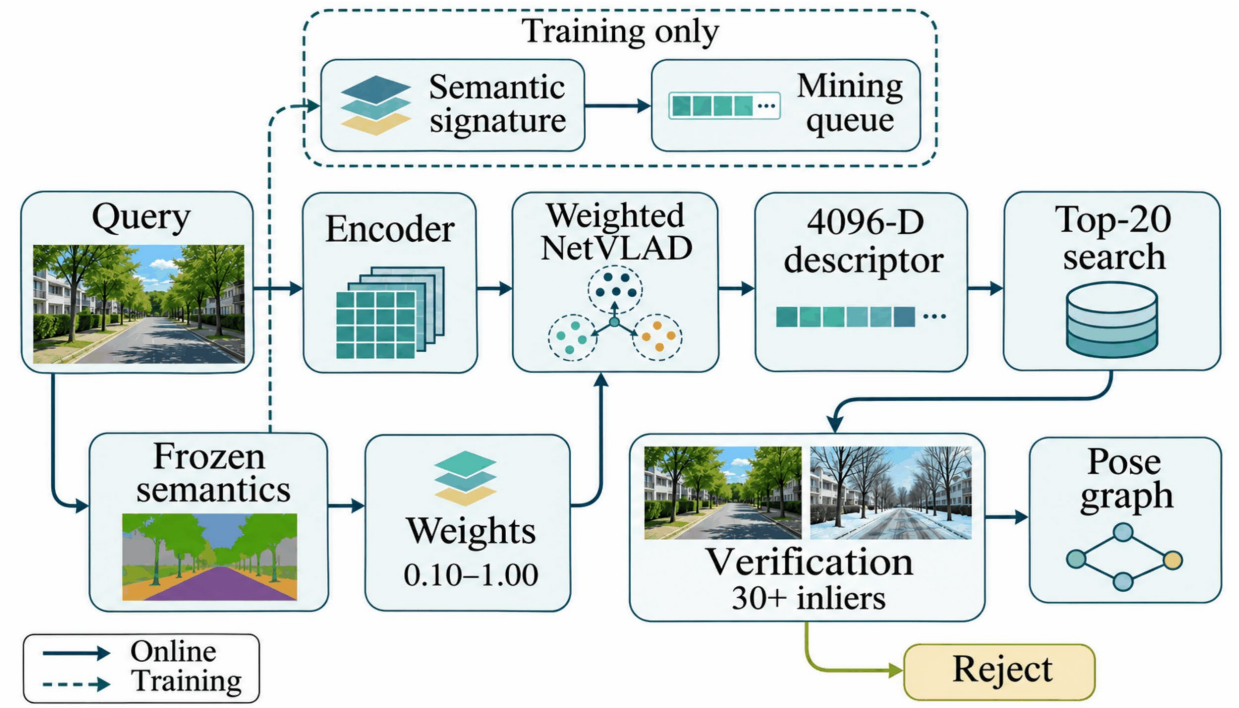


Figure 1. Confidence weighted NetVLAD architecture

3.2. Geographically Guarded Semantic Hard Negatives

Each training observation contains an image identifier, a traversal identifier, an approximate position, and a seasonal condition label. Positive pairs are selected within a nominal 10 m place tolerance and screened for adequate view overlap. Negative candidates lie beyond 30 m, leaving an intermediate region unused. This exclusion region provides a conservative spatial guard whose adequacy depends on positioning uncertainty. Geometric checks and route metadata support the assessment of correspondence, while observations from overlapping route sections with unresolved alignment receive an ambiguity flag. Pair validity is established before descriptor similarity influences candidate selection. The term safe denotes eligibility under the stated geographic and ambiguity filters. Residual positioning error and incomplete overlap information remain possible sources of pair-label uncertainty.

A semantic signature is constructed by averaging category probabilities within a coarse spatial grid and concatenating the resulting distributions. Normalization allows cosine similarity to describe resemblance in category composition and spatial layout. A memory queue stores recent descriptors, semantic signatures, and pair validity metadata. Within the geographically admissible set, candidate ranking combines semantic resemblance with descriptor proximity through the equation below: Grid resolution, retained category vocabulary, and confidence aggregation are signature-construction parameters. A reproducible implementation must publish these settings together with the normalization convention and semantic screening threshold.

$$h(q, n) = \alpha s(q, n) - (1 - \alpha) d(q, n), n \in \mathcal{N}_{safe}(q) \quad (2)$$

In this equation, q denotes an anchor observation, n is a candidate negative, $s(q, n)$ measures semantic cosine similarity, and $d(q, n)$ represents descriptor distance on a fixed bounded scale. The coefficient α controls the relative influence of semantic resemblance and descriptor proximity. The admissible set $\mathcal{N}_{safe}(q)$ contains candidates retained after geographic and ambiguity filtering. Fixed scaling preserves the interpretation of α as training progresses. A semantic threshold first removes scenes with weak category resemblance; the remaining candidates are ranked by $h(q, n)$. Sampling within a difficulty band preserves variation among selected negatives, particularly when multiple neighboring database images depict the same street segment. The mining distance is a bounded ranking quantity, whereas the training objective uses squared Euclidean distance between normalized descriptors. The margin belongs to the objective's distance scale; the semantic coefficient belongs to the mining score's fixed scale.

The curriculum uses the middle difficulty band during the first five epochs, admits candidates from the upper half through epoch fifteen, and includes the highest ranked candidates thereafter. The queue refreshes every epoch and records the encoder version associated with each descriptor. Entries older than two refreshes are removed to limit stale ranking decisions. Each anchor receives one verified positive and up to eight safe negatives. When fewer candidates are available, all admissible negatives are retained and their contributions are averaged within the anchor. The geographic guard remains fixed throughout this process. Route segments define the sampling diversity unit, allowing negative selection to cover distinct locations while controlling the influence of adjacent frames. Queue age controls descriptor

freshness, while geographic and overlap metadata control pair eligibility. Refreshed descriptors retain the same validity restrictions, including exclusion of known positives and unresolved correspondences. Difficulty bands refer to ranks within the admissible candidate pool. Exact percentile boundaries and tie handling require development-set specification; the epoch schedule alone determines when each configured band becomes available.

Figure 2 illustrates the sequence of geographic filtering, ambiguity rejection, semantic screening, descriptor ranking, curriculum sampling, and queue refresh. Ambiguous pairs remain outside both supervision sets until their validity can be established. Training logs record candidate availability, rejected pair counts, difficulty distributions, and the proportion of anchors receiving a complete negative set. These records distinguish changes in sample difficulty from changes in the quantity of available supervision.

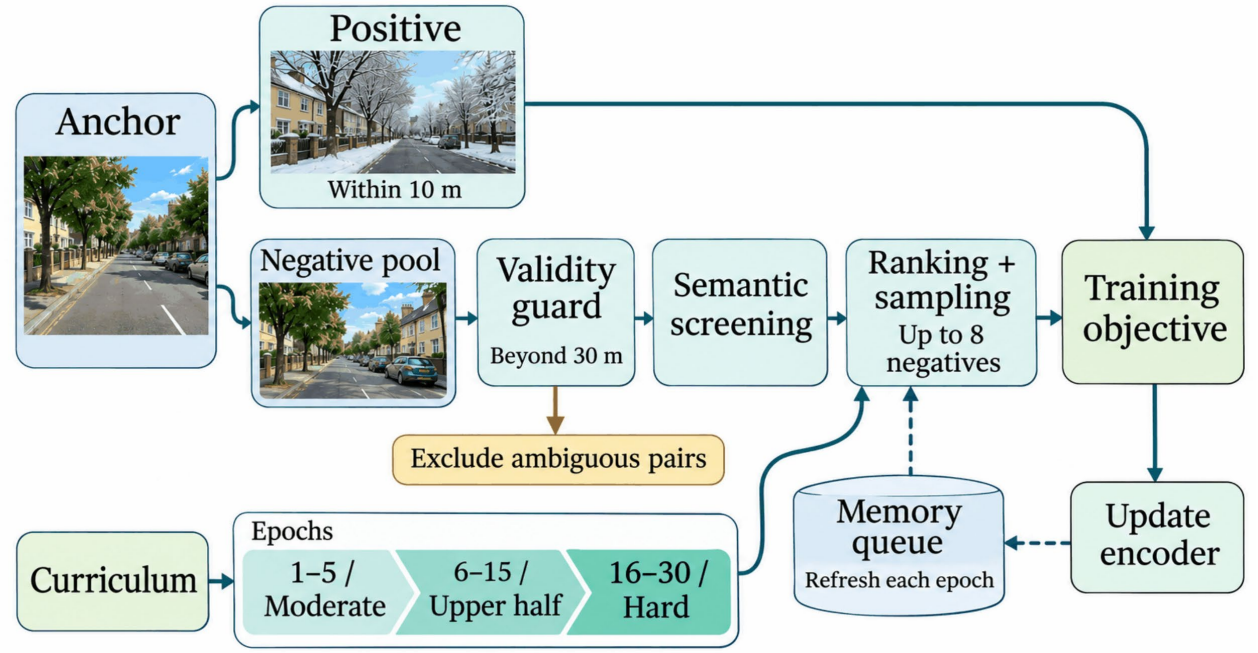


Figure 2. Guarded semantic hard negative mining

3.3. Objective and Loop Closure Interface

Descriptor learning combines margin ranking with a seasonal consistency penalty. The ranking component compares each anchor with its verified positive and selected negatives. Positive observations are drawn from a different traversal whenever suitable supervision is available. Distances are computed as squared Euclidean distances between unit normalized descriptors, and the margin is fixed at 0.20 using the development partition. To preserve equal anchor influence when candidate availability varies, the ranking equation averages negative contributions within each anchor before averaging across the batch:

$$\mathcal{L}_{\text{rank}} = \frac{1}{B}\sum_{q\subset\mathcal{B}} \frac{1}{|\mathcal{N}_q|}\sum_{n\subset\mathcal{N}_q} \max\left(0, m + d\left(q, p_q\right) - d(q, n)\right) \tag{3}$$

Here, $\mathcal{B}$ is a batch containing B anchors, p_q denotes the verified positive associated with anchor q, and $\mathcal{N}_q$ contains its selected safe negatives. Anchors enter the ranking batch when at least one admissible negative is available. The margin m specifies the desired separation between positive and negative distances. Consequently, a negative contributes to optimization when its distance fails to exceed the positive distance by the required margin. This formulation directs training toward unresolved ranking errors while maintaining comparable contributions from anchors with different numbers of safe candidates. Anchors with no admissible negative are excluded from the ranking batch, and the batch average is defined over the remaining eligible anchors. This avoids treating an empty negative set as a zero-loss training example.

The auxiliary term encourages consistency between verified cross-seasonal positive descriptors. Geographic and overlap supervision determine these pairs, preserving discrimination between distinct streets with similar semantic layouts. A coefficient λ of 0.10 gives seasonal consistency a supporting role. Combining the ranking and consistency terms yields the training objective expressed by the following equation: The consistency term reinforces cross-seasonal positive attraction already present in ranking, while the negative terms preserve separation from distractors. Its coefficient controls this additional emphasis and requires sensitivity analysis before transfer to other route distributions.

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \frac{\lambda}{|\mathcal{P}|} \sum_{(q,p)\subset\mathcal{P}} \left\|\mathbf{z}_q - \mathbf{z}_p\right\|_2^2 \tag{4}$$

The set $\mathcal{P}$ contains verified cross-seasonal positive pairs, and $\mathbf{z}_q$ and $\mathbf{z}_p$ are their unit normalized descriptors. The ranking term maintains separation from geographically distinct distractors, while the consistency term reduces variation between corresponding seasonal observations. Their relative weighting determines how strongly cross-seasonal attraction influences the discriminative representation.

Optimization uses thirty epochs, batches of sixteen anchors, and an initial learning rate of 0.0001, reduced after epochs fifteen and twenty-five. The encoder, aggregation layer, and projection are trained jointly, while the semantic predictor remains fixed. Matched baselines receive the same initialization, image exposure, augmentation schedule, and descriptor dimensionality. Geometric augmentation preserves plausible view overlap. Independent color and contrast perturbations encourage positive pairs to retain correspondence under appearance variation. The primary comparison uses a common augmentation policy across configurations, allowing sampling and structural weighting effects to be assessed under consistent image exposure. Gradient clipping controls occasional large updates produced by difficult batches.

During deployment, the normalized descriptor retrieves twenty candidate keyframes. A temporal exclusion window removes observations close to the current frame within the same traversal, and a descriptor threshold screens the remaining proposals before local correspondence verification. Geometric acceptance requires at least thirty inliers with adequate spatial distribution. This distribution check limits the influence of matches concentrated on a small repeated pattern. Local correspondence geometry supplies the transformation estimate, and accepted candidates enter the mapping back end with their verification statistics. Rejected proposals remain in the diagnostic log. Descriptor thresholds, database construction rules, and verifier settings are fixed before evaluation, enabling retrieval behavior and geometric acceptance to be assessed under a consistent operating policy. The inlier count is one acceptance requirement. Matcher configuration, transformation model, residual tolerance, spatial coverage criterion, and robust-estimation settings are additional verifier dependencies that must be fixed before evaluating accepted constraints. Temporal exclusion applies to neighboring observations within the same traversal. Candidate eligibility across separate traversals follows geographic and overlap checks, so temporal proximity is interpreted within a traversal-specific sequence.

4. Experimental Evaluation and Analysis

4.1. Protocol and Seasonal Retrieval Behavior

The evaluation follows a controlled seasonal retrieval protocol with separate assessments of ranking, component effects, computational behavior, and loop acceptance. The 4Seasons benchmark jointly evaluates visual odometry, global place recognition, and map-based localization across seasonal, weather, and illumination changes [24]. The evaluation configuration contains four seasonal transition groups, each comprising 1500 query locations and a database of 10000 reference keyframes. Training, development, and evaluation receive sixty, twenty, and twenty percent of geographically separated route segments, respectively, with a 100 m boundary buffer. A valid retrieval lies within a 10 m place tolerance and satisfies view overlap requirements. Closed set recall uses queries with a valid reference revisit, while open set acceptance uses queries whose reference database contains no corresponding place. A similar separation of processing stages appears in opposite-viewpoint recognition, where coarse retrieval is followed by semantically filtered keypoint matching [25]. The matched configurations share a visual encoder, a 4096-dimensional output, image augmentation, and optimization exposure. Development routes determine mining coefficients and acceptance thresholds before evaluation. This separation assigns model selection and final assessment to different observations and makes the spatial tolerance, database density, and candidate budget explicit experimental controls.

Figure 3(a) compares top ranked recall across summer to winter, autumn to winter, spring to winter, and winter to summer transitions. The baseline values are 79.1%, 76.8%, 73.5%, and 70.2%, while the complete configuration reaches 88.4%, 86.1%, 83.7%, and 81.8%. The absolute improvements are therefore 9.3, 9.3, 10.2, and 11.6 percentage points. The winter to summer case has the largest improvement and the lowest final recall, indicating that the strongest gain can coexist with substantial residual difficulty. The baseline range across transitions is 8.9 points; the complete configuration reduces this range to 6.6 points. This narrowing indicates a more uniform response across the four seasonal transitions. The smaller range is consistent with the intended reduction in transition sensitivity through structural weighting and lookalike negative selection. Seasonal asymmetry could also arise from differences in visible landmarks, camera orientation, or reference density. Those variables require separate recording before the observed ordering can be attributed to a particular source of appearance change.

Figure 3(b) examines mean recall at retrieval depths of 1, 5, 10, 20, and 50, where the complete model reaches 85.0%, 91.2%, 93.8%, 95.1%, and 96.4%. Increasing depth from one to five recovers 6.2 percentage points, whereas increasing it from twenty to fifty recovers 1.3 points. The declining marginal return places the twenty-candidate budget in a region

where most recoverable places are already represented. Database composition remains relevant: multiple nearby keyframes can occupy several positions while representing one physical place. Spatiotemporal attention methods explicitly model persistence and change across neighboring frames when constructing sequence descriptors [26]. Figure 3(c) presents the complementary precision response at recall levels of 20%, 40%, 60%, 80%, and 90%. The complete configuration retains precision values of 99.8%, 99.5%, 98.9%, 96.8%, and 90.4%. Its 6.4-point precision decrease between the last two operating points indicates an increasingly difficult distractor population. Expanding retrieval therefore creates two separate effects: recovery of correct candidates and a greater verification burden. A runtime assessment must measure the local matching work associated with that burden.

Figure 3(d) retains the full method by transition matrix, including structural weighting and guarded semantic mining. Weighting gives recalls of 82.0%, 80.0%, 77.0%, and 73.0%, while mining gives 85.0%, 83.0%, 80.0%, and 76.8%. Mining exceeds weighting by 3.0, 3.0, 3.0, and 3.8 percentage points, exposing a distinct contribution from the training pair distribution. The complete model then improves on mining by 3.4, 3.1, 3.7, and 5.0 points. This last increment combines structural weighting and curriculum effects and therefore requires the finer component comparison that follows. The largest increment again occurs in winter to summer retrieval. Because all configurations retain the same descriptor dimension, the matrix isolates differences in feature use and supervision at a common output capacity. The heatmap exposes transition specific weaknesses alongside the aggregate mean. Per query rankings should accompany these aggregate cells so that gains from rescuing previously missed locations can be separated from changes among queries already recognized correctly.

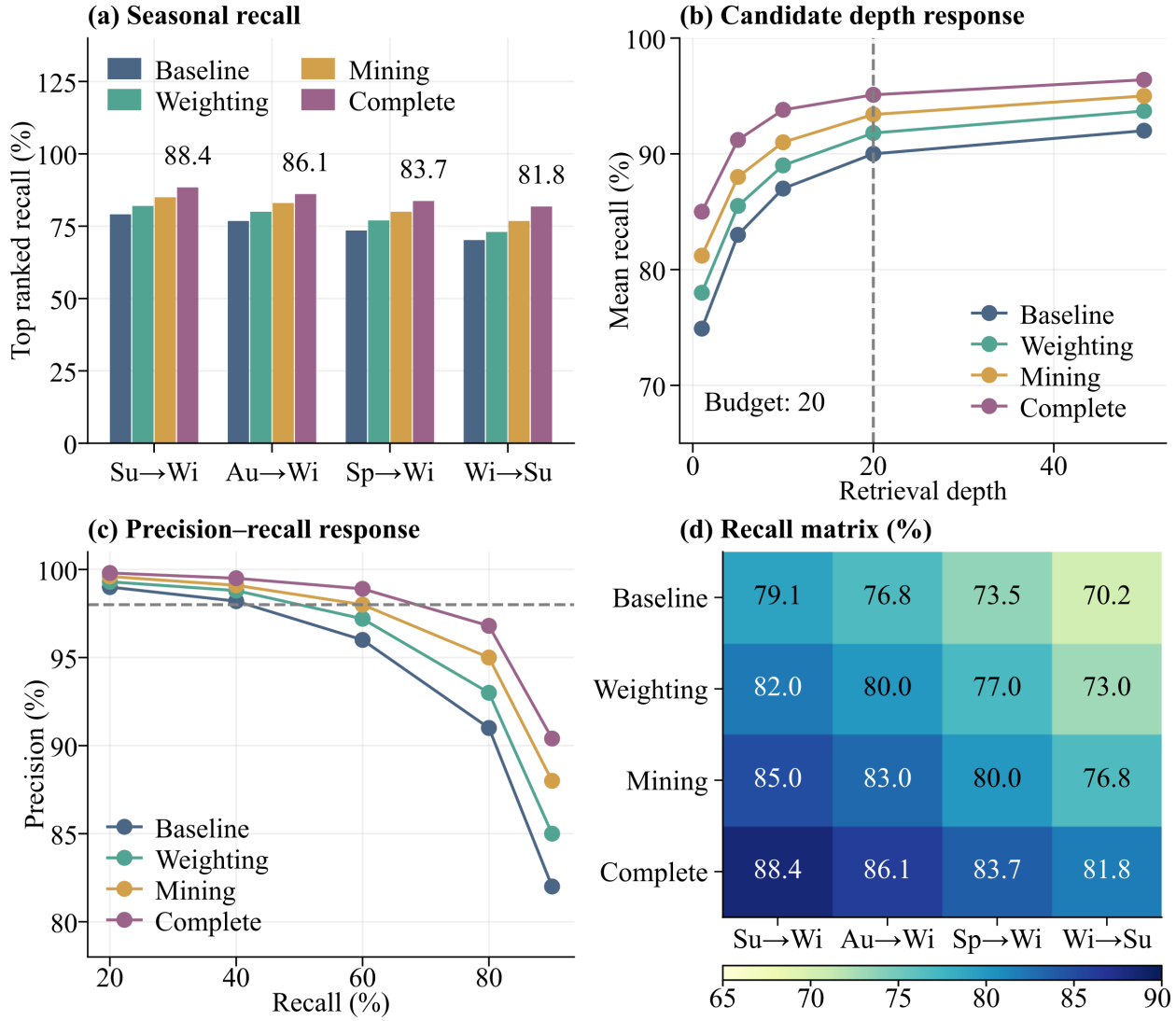


Figure 3. Seasonal retrieval performance: (a) grouped recall bars; (b) recall versus retrieval depth; (c) precision versus recall; (d) transition by method heatmap

Geographic tolerance deserves a separate sensitivity check because the same ranked list can receive different labels when the acceptance radius changes. Geometric usability requires additional checks, as demonstrated by learned homography estimation for verifying the spatial consistency of local feature matches [27]. The specified 10 m positive tolerance and 30 m negative threshold leave a 20 m unused region for uncertain pairs. A stricter positive tolerance would

remove some currently valid references; a looser one could include views with weak geometric overlap. The comparison fixes both settings so that their influence remains separate from the method comparison. A physical evaluation should report results at several tolerances, preserve the original query set, and identify queries whose correctness changes near route boundaries. Frame spacing and heading should also be retained because densely sampled views can inflate candidate coverage. The winter to summer transition is especially useful for this inspection: a query may retrieve the correct road segment yet favor a neighboring viewpoint with substantially different visible structure. Reviewing those boundary cases would distinguish seasonal descriptor errors from label uncertainty and insufficient overlap.

4.2. Component Effects and Decision Sensitivity

Joint descriptor and pair-classifier training provides a framework for studying the interaction between global retrieval and candidate re-ranking [28]. Table 1 therefore keeps the 4096-dimensional representation fixed across five configurations. Mean top ranked recall increases from 74.9% for the baseline to 78.0% with structural weighting, 81.2% with guarded semantic mining, and 83.1% when both are combined without the curriculum. The complete configuration reaches 85.0%, assigning a 1.9-point increment to the curriculum under this specific control. The weighting and mining increments of 3.1 and 6.3 points have different interpretations: the former changes spatial evidence within a descriptor, while the latter changes the mistakes emphasized during optimization. Their combined effect can interact because weighting also changes which candidates appear difficult. Verification precision rises from 98.1% to 99.1%, while median latency rises from 28.0 to 35.0 ms. Precision is conditional on accepted proposals, so the table should be read alongside retrieval coverage. The latency difference reflects the configured online semantic branch and search workload; offline queue maintenance belongs to training resource accounting. Each ablation increment is conditional on the components retained in its comparison. Pooling changes descriptor geometry and therefore the mined candidates, so the separate gains should not be added as independent estimates of component effect.

Table 1. Component comparison with matched descriptor dimensions

Configuration	Recall (%)	Precision (%)	Latency (ms)
NetVLAD baseline	74.9	98.1	28.0
Structural weighting	78.0	98.3	33.0
Guarded semantic mining	81.2	98.6	30.0
Combined without curriculum	83.1	98.8	34.0
Complete model	85.0	99.1	35.0

Figure 4(a) traces development recall through epochs 1, 5, 10, 15, 20, and 30. All four configurations begin at 40.0%; their final values are 74.9%, 78.0%, 81.2%, and 85.0%. At epoch fifteen, the complete model reaches 80.0%, exceeding the mining configuration by 3.0 points; by epoch thirty, the difference is 3.8 points. The continued separation after admission of the hardest candidates is consistent with the intended role of the curriculum, although a controlled schedule comparison is required to identify causality. The final complete model increment from epoch twenty to thirty is 2.0 points, compared with 1.9 points for the baseline, indicating similar late-stage gains across the two training trajectories. Figure 4(b) varies semantic score weight over 0.10, 0.25, 0.40, 0.55, 0.70, and 0.85, producing recalls of 80.2%, 82.6%, 84.4%, 85.0%, 84.1%, and 81.5%. The intermediate maximum supports retaining both semantic resemblance and descriptor difficulty in candidate ranking. Excessive semantic emphasis can repeatedly select similar category layouts while reducing the diversity of informative descriptor errors.

Figure 4(c) compares negative distance distributions centered near 0.86 for random selection and 0.55 for guarded semantic selection. The diagnostic marker at 0.60 identifies a region containing closer distractors. The shift toward smaller distances indicates that mining increases the proportion of selected candidates that challenge the current ranking. The overlap between distributions remains important because semantic screening retains variation within its selected population. A policy selecting only the closest observations could concentrate supervision on a narrow set of repeated structures. Noise-robust deep metric learning uses clean-sample selection and subgroup-based positive-pair construction to improve the reliability of training pairs [29]. Here the geographic guard and ambiguity filter establish admissibility before distance ranking. An implementation should record the fraction of selected pairs below the diagnostic marker together with geographic distance and overlap audit results. Those records would reveal whether a curriculum stage increases useful confusion, stale queue effects, or unresolved positive correspondence. The histogram characterizes selection pressure, and pair review evaluates its consistency with place identity.

Figure 4(d) places 120 safe negatives in semantic similarity and descriptor distance coordinates, with curriculum stage shown by color. Semantic resemblance spans approximately 0.30 to 0.98, while the distance population follows a decreasing trend with substantial residual spread. At a similar semantic score, different candidates can occupy markedly different descriptor distances. This spread distinguishes the roles of semantic screening and descriptor ranking: screening

restricts scene composition, while ranking identifies errors in the learned representation. The early, middle, and late groups illustrate progressively stronger semantic screening while preserving variation within each group. A candidate with high semantic similarity and large descriptor distance contributes little immediate margin pressure, whereas a close candidate with uncertain geographic validity requires exclusion. Queue version metadata becomes relevant near stage boundaries because an old descriptor may misrepresent current difficulty. Logging the age and encoder version of each selected entry would allow stage effects to be separated from delayed candidate updates.

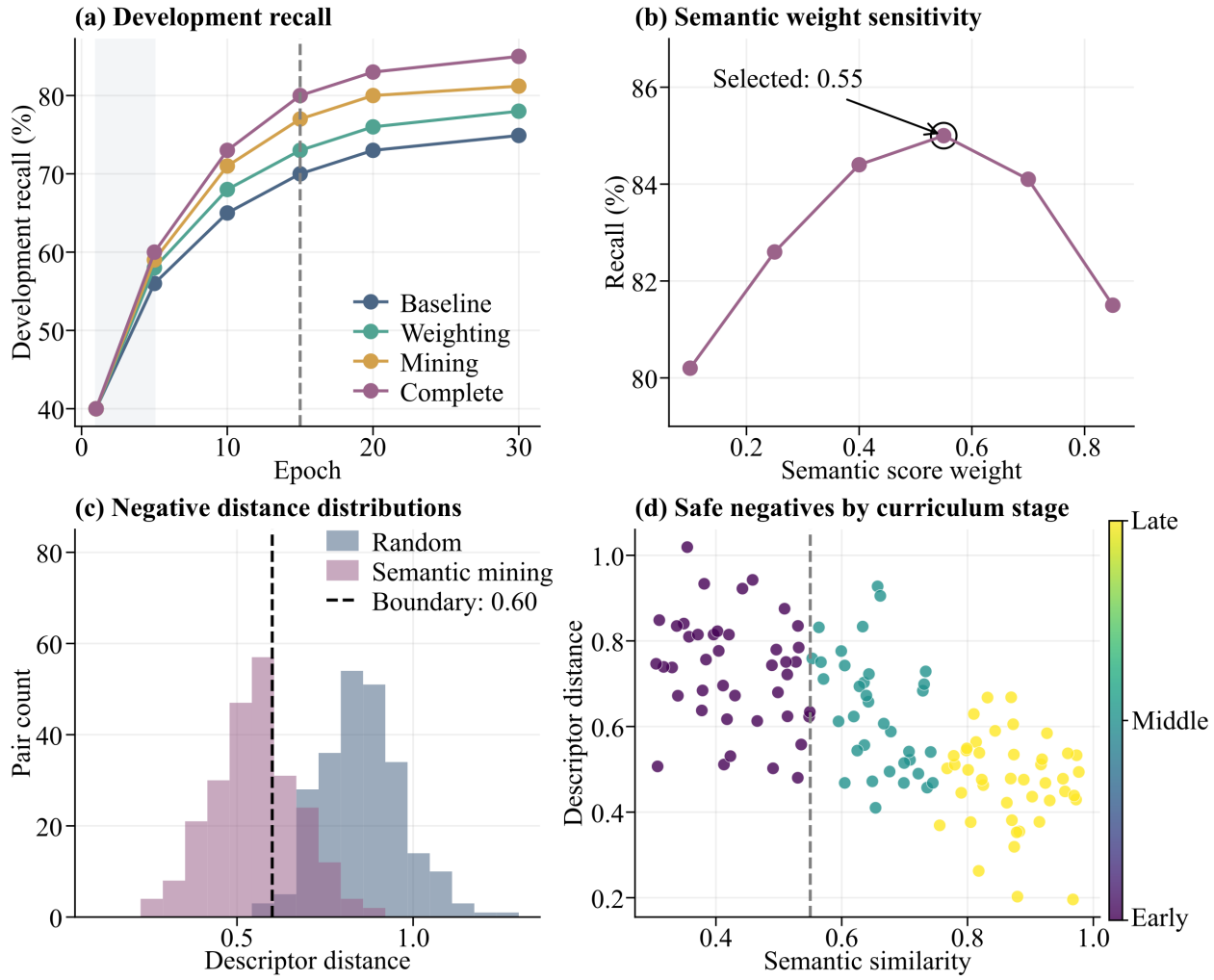


Figure 4. Training and mining diagnostics: (a) development recall curves; (b) semantic weight response; (c) negative distance distributions; (d) semantic resemblance and descriptor distance scatter

Figure 5(a) displays twelve route block values for each configuration. The block centers are 74.9%, 78.0%, 81.2%, and 85.0%, with an equal spread of six percentage points. The equal spread facilitates comparison of the central shift at a common level of route block variability. Figure 5(b) displays gains of 9.3, 9.3, 10.2, 11.6, and 10.1 points for the seasonal transitions and their mean. Its descriptive interval half widths are 1.3, 1.6, 1.4, 1.8, and 1.1 points. Correlation-aware sequential embeddings exploit dependencies between consecutive frames to support sequence retrieval and frame-level matching [30]. For the present variability analysis, paired model results should be retained within independently collected route blocks. The largest span belongs to winter to summer retrieval, directing scrutiny toward the transition with the strongest gain and greatest residual difficulty. Independent traversals would determine whether this apparent variability persists across environmental observations. The displayed spans describe the configured variability range and are not inferential confidence intervals. Statistical significance and population coverage would require independent route observations and a stated uncertainty-estimation procedure.

Figure 5(c) stratifies complete model recall by semantic confidence centers of 0.30, 0.45, 0.60, 0.75, and 0.90. Recall rises from 68.0% through 76.2%, 82.3%, and 86.9% to 90.1%, a 22.1-point difference between the extreme bins. This gradient identifies low confidence queries as a priority for inspection. It could reflect segmentation ambiguity, poor image quality, or reduced structural visibility, so confidence itself should be tested against observed prediction correctness. The increase from 0.30 to 0.45 is 8.2 points, larger than the 3.2-point increase between the final two bins. Such a pattern would favor targeted intervention on uncertain scenes over globally increasing semantic weight. Bin populations must accompany the analysis because a sparse low confidence bin can produce an unstable recall estimate.

The same stratification should be applied to the unweighted baseline to establish whether confidence marks general visual difficulty or a specific sensitivity introduced by semantic pooling.

Figure 5(d) separates the perturbation profile into snow, foliage, glare, occlusion, viewpoint change, and repeated facades, with scores of 82, 87, 80, 78, 84, and 76. Repeated facades are 11 points below foliage and form the weakest category. That ordering exposes a residual distinction between suppressing seasonal appearance and preserving unique location cues. Broad building categories can remain stable across seasons while repeated windows and street layouts still produce descriptor confusion. Occlusion is the second weakest category at 78, indicating a different failure mechanism in which useful structure is hidden. A follow up evaluation should annotate visible structural area and repeated pattern density separately, then compare the mining policy under matched visibility. Facade errors could motivate finer spatial signatures, while occlusion errors could require local evidence selection or multiple observations. Testing those responses independently would show whether the mining candidate distribution contains the information needed to address each failure type. The radial values are condition-specific descriptive scores. Their normalization and category construction require a stated protocol before they can be interpreted as calibrated probabilities or compared quantitatively across evaluation studies.

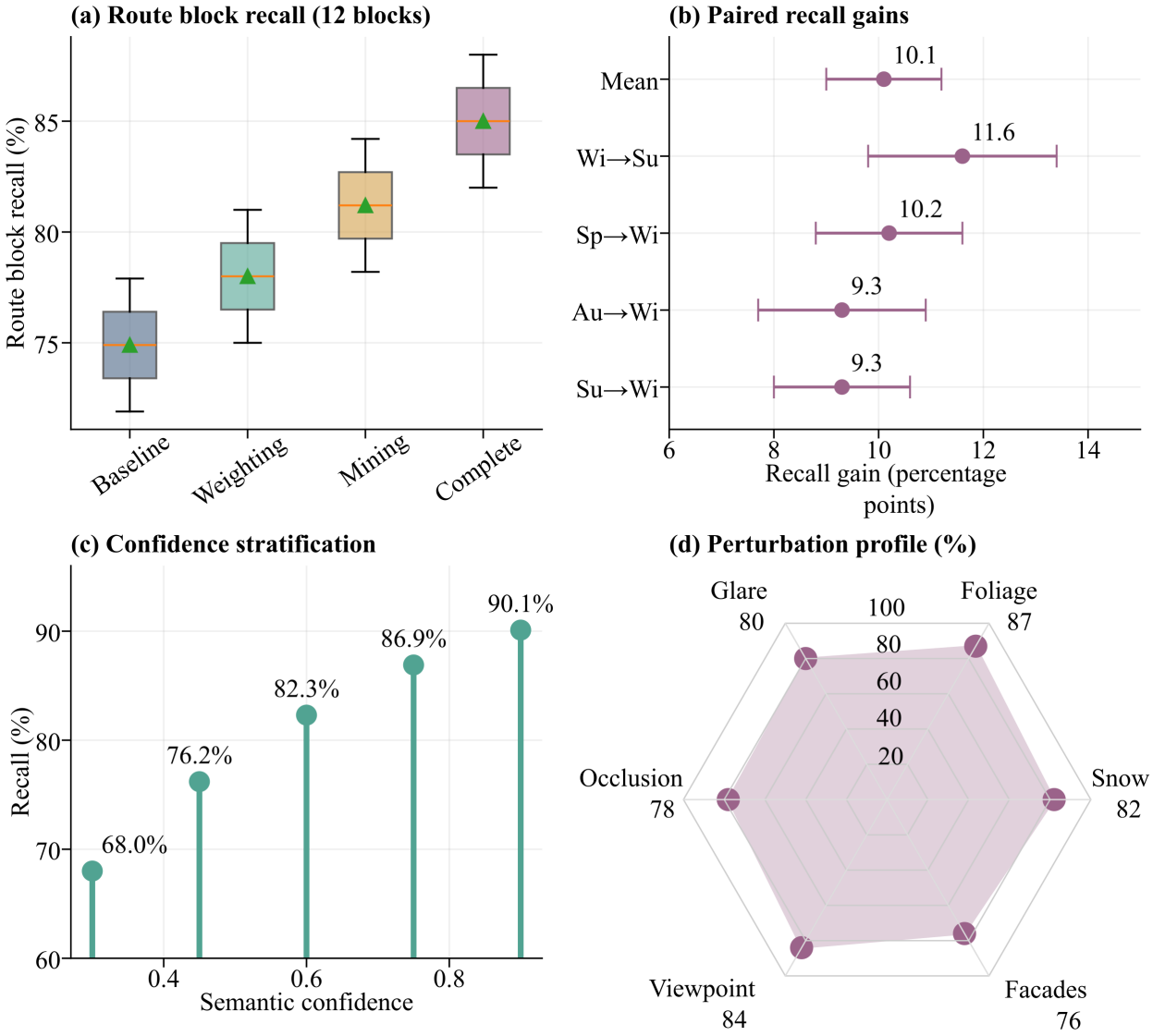


Figure 5. Variability and robustness: (a) route block box plots; (b) gain interval plot; (c) confidence bin lollipop plot; (d) perturbation radial profile

4.3. Resource Behavior and Failure Analysis

The resource protocol separates visual encoding, semantic inference, aggregation, database search, and local verification. Deployment studies of compact and quantized place recognition networks jointly evaluate retrieval recall, memory consumption, and latency [31]. The 35.0 ms budget assigns 18.0 ms to visual encoding, 7.0 ms to semantic processing, 4.0 ms to aggregation and projection, and 6.0 ms to search and verification. These allocations identify the contribution of each processing stage to the total budget. Image acquisition and map optimization require additional

system level accounting, while the training queue belongs to offline resource use. Deployment assessment should report hardware, software versions, batch size, index settings, warmup procedure, and both median and upper tail latency. Difficult scenes may increase verification work even when feature extraction time remains stable. Candidate count and rejection stage should therefore accompany per query timing. The semantic branch must remain within the deployed model footprint because it generates the spatial weights used by the proposed online descriptor. The latency decomposition specifies a front-end resource budget. A measured throughput claim requires hardware identification, accelerator synchronization, repeated execution traces, and accounting for acquisition and mapping-back-end work.

Figure 6(a) shows complete model median delays of 29.0, 31.0, 35.0, 43.0, and 56.0 ms for databases containing 1000, 5000, 10000, 25000, and 50000 keyframes. The corresponding descriptive half ranges are 2, 3, 4, 6, and 8 ms. Increasing the database fivefold from 10000 to 50000 entries adds 21.0 ms to the median, while the range widens from 4 to 8 ms. This difference directs attention to search and verification variability as the map grows. The median crosses the 50 ms budget between the last two database sizes. At 25000 entries the upper displayed value is 49 ms, leaving little allowance for additional system overhead. Index selection, candidate pruning, and the number of local matching attempts should therefore be profiled separately before a processing deadline is assigned. The curve identifies the database sizes at which search and verification demand further profiling.

Figure 6(b) gives single precision descriptor payloads of 16.4, 81.9, 163.8, 409.6, and 819.2 MB for the same database sizes, using decimal megabytes. A 4096-component descriptor occupies 16384 bytes, which explains the proportional increase with keyframe count. The half precision alternative occupies 8.2, 41.0, 81.9, 204.8, and 409.6 MB. The saving at 50000 keyframes is therefore 409.6 MB in descriptor payload. Search index structures, cached images, local features, and network parameters require separate memory allocation. The filled regions make the two payload policies comparable across the entire database range. Their separation is determined by storage format, while any associated retrieval change requires a precision sensitivity experiment. Closely ranked candidates are a useful inspection set because small representation perturbations can change their ordering even when overall vector reconstruction error is modest. Storage values use decimal megabytes and count descriptor payload only. Index structures, images, local features, and model parameters require additional memory; half-precision payload reduction must be evaluated separately from its effect on retrieval ordering.

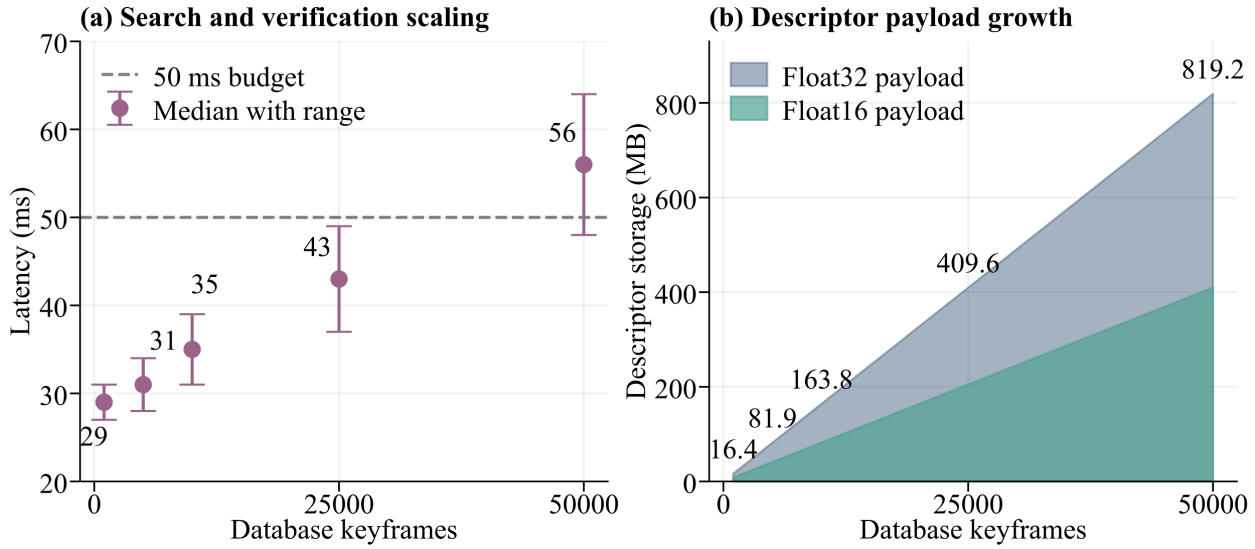


Figure 6. Deployment scaling: (a) database size versus latency with a 50 ms budget; (b) descriptor payload storage at single and half precision

Database growth also requires evaluation of the keyframe selection policy. Storing neighboring observations densely can improve view overlap but increase search work and duplicate candidate proposals. A controlled resource test should vary reference spacing while keeping route coverage fixed, recording both retrievals recall and the number of distinct places among the twenty candidates. This would separate useful geographic coverage from repeated views of one location. Combining keyframe spacing with descriptor precision creates a second tradeoff: sparse high precision references may preserve discrimination, while dense compressed references may preserve viewpoint coverage. Per query error inspection can identify which factor limits a particular route. Routes with sharp turns may benefit from additional viewpoints, whereas long uniform facades can accumulate redundant entries. The storage curve provides a reproducible accounting basis for those alternatives, and the latency curve identifies the database region in which candidate handling becomes a potential bottleneck.

Semantic query rejection evaluates whether an image contains sufficient place-representative context to support a recognition decision [32]. Figure 7(a) evaluates thresholds of 0.50, 0.60, 0.70, 0.80, 0.90, and 0.95, with false acceptance

rates of 2.8%, 1.7%, 0.9%, 0.4%, 0.2%, and 0.1%. Missed loop rates at the same settings are 4.0%, 6.0%, 9.0%, 15.0%, 27.0%, and 41.0%. Moving from the 0.70 threshold to 0.80 reduces false acceptance by 0.5 points but increases missed loops by 6.0 points. Increasing the threshold from 0.90 to 0.95 reduces false acceptance by 0.1 percentage point while increasing the missed loop rate by 14.0 percentage points. These changes indicate a substantial reduction in loop coverage at high acceptance thresholds. Threshold selection should use the expected frequency of valid revisits and the mapping consequences of an incorrect constraint. Both query populations require independently checked ground truth and separate rate denominators. False acceptance is counted per query without a valid database revisit, after geometric verification. Missed loops are counted per query with a valid revisit; the two percentages therefore use different query populations and must be interpreted separately.

Figure 7(b) divides 200 inspected errors into repeated facades, vegetation occlusion, snow concealment, semantic uncertainty, and viewpoint mismatch, with counts of 58, 44, 39, 33, and 26. Their respective shares are 29.0%, 22.0%, 19.5%, 16.5%, and 13.0%. Facade repetition and vegetation occlusion jointly account for 51.0% of the reviewed error set, identifying two distinct priorities for error review. The former preserve's visible structure but weakens uniqueness; the latter removes or mask's informative structure. Loop detection in agricultural environments combines local feature search with stereo geometric refinement and relative pose estimation [33]. The proposed verification interface checks inlier distribution alongside inlier count. Inspecting accepted facade errors would establish whether incorrect matches cluster on one repeated region or span a larger ambiguous pattern. Snow and semantic uncertainty require additional review of the confidence map, since a structurally useful surface may be concealed or assigned an unreliable category.

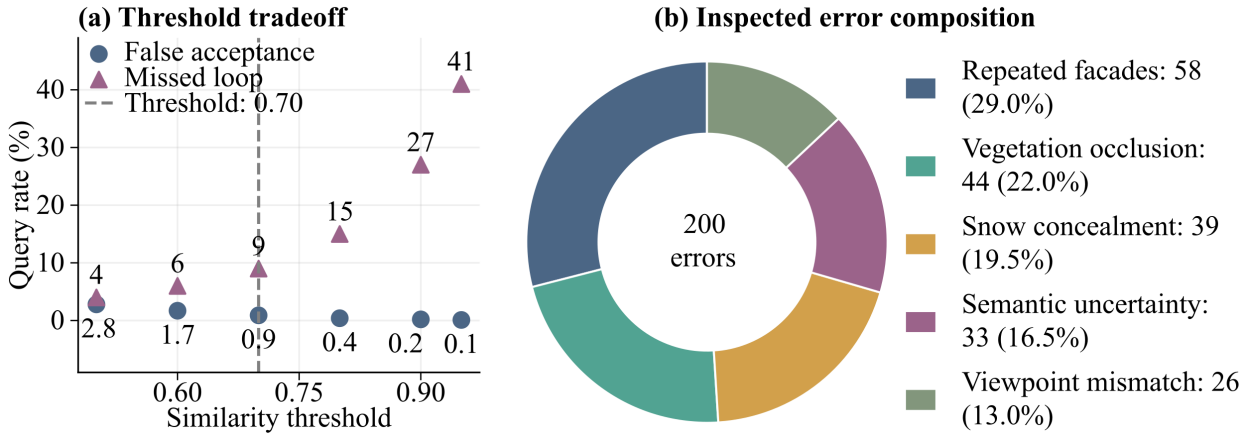


Figure 7. Acceptance and failure behavior: (a) threshold tradeoff scatter with operating point annotations; (b) inspected error composition donut with counts and percentages

Multimodal benchmarking has systematically compared camera, LiDAR, semantic, and text inputs on the Oxford RobotCar and NCLT datasets with publicly released implementation code [34]. Each reviewed query should retain its ranked candidates, semantic confidence, negative validity metadata, verification outcome, and processing delay. The error categories above should be assigned through a documented coding procedure, including a rule for scenes containing multiple causes. A high confidence semantic error is particularly informative: it can preserve an unsuitable pooling weight while directing the mining queue toward misleading category layouts. Comparing confidence calibration across snow, low light, and foliage conditions would expose that failure pathway. A separate route dominated by vegetation would test whether reduced vegetation weights remove useful spatial evidence when buildings are scarce. For accepted candidates, local correspondence geometry and the subsequent pose graph residual should be recorded together. Those records would distinguish a correct place proposal with weak geometric support from an incorrect proposal that survives verification, enabling targeted changes to weighting, sampling, and acceptance criteria.

5. Conclusion

This paper developed a NetVLAD based loop closure front end that connects confidence weighted structural aggregation with geographically guarded semantic hard negative mining. The design uses semantic resemblance to identify informative distractors and independently validated supervision to establish place identity. A gradual mining schedule controls the exposure to difficult candidates, and the deployment interface preserves geometric verification as a separate acceptance stage. The evaluation establishes a coherent set of comparisons through which these mechanisms can be examined, including seasonal ranking, component effects, operating thresholds, resource behavior, and recognizable failure conditions. Application-level reliability requires independently recorded traversals, measured execution traces, and verification of accepted transformations within a complete mapping back end.

The scope of the evaluation is defined by its route configuration, seasonal transition coverage, and fixed acceptance settings. Broader geographic coverage and independent seasonal traversals remain necessary for assessing transferability.

Coarse semantic signatures may fail to distinguish repeated structures, and confidence weighting depends on the calibration of a predictor exposed to changing environmental conditions. Geographic guards can also retain mislabeled pairs when positioning uncertainty or viewpoint ambiguity exceeds the assumptions used to construct them. The reliability of a loop transformation and its effect on a complete mapping trajectory require further geometric and system level validation. Priority checks concern negative-pair validity, semantic confidence calibration, and geometric acceptance reliability. These checks respectively address supervision errors, misleading spatial weights, and incorrect map constraints, providing a direct basis for subsequent validation.

Future work should implement the training pipeline on geographically separated seasonal routes and publish the observations needed to reproduce both sampling and evaluation. Particular attention should be given to audited negative validity, semantic calibration under domain shift, and uncertainty estimates based on independent traversals. Integration with a complete mapping back end would permit direct assessment of accepted constraints and trajectory consistency. Adaptive spatial guards, richer representations of structural layout, and calibrated rejection policies offer further directions for improving the connection between persistent visual evidence and dependable long term loop closure.

6. Declarations

6.1 Author Contributions

Kamil Wysocki contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Adrian Ostrowski contributes to validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

6.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

6.3 Institutional Review Board Statement

Not applicable.

6.4 Informed Consent Statement

Not applicable.

6.5 Date Availability

Not applicable.

6.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

7. References

[1] Schubert, S., Neubert, P., Garg, S., Milford, M., & Fischer, T. (2024). Visual place recognition: A tutorial [Tutorial]. IEEE Robotics & Automation Magazine, 31(3), 139–153. https://doi.org/10.1109/mra.2023.3310859

[2] Jiang, H., Piao, S., Yu, H., Li, W., & Yu, L. (2024). Robust visual place recognition for severe appearance changes. IEEE Robotics and Automation Letters, 9(5), 4289–4296. https://doi.org/10.1109/lra.2024.3376967

[3] Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K. M., Scherer, S., Krishna, M., & Garg, S. (2024). AnyLoc: Towards universal visual place recognition. IEEE Robotics and Automation Letters, 9(2), 1286–1293. https://doi.org/10.1109/lra.2023.3343602

[4] Claxton, O., Malone, C., Carson, H., Ford, J. J., Bolton, G., Shames, I., & Milford, M. (2024). Improving visual place recognition based robot navigation by verifying localization estimates. IEEE Robotics and Automation Letters, 9(12), 11098–11105.

https://doi.org/10.1109/lra.2024.3483045

[5] Uggi, A., & Channappayya, S. S. (2024). MS-NetVLAD: Multi-scale NetVLAD for visual place recognition. IEEE Signal Processing Letters, 31, 1855–1859. https://doi.org/10.1109/lsp.2024.3425279

[6] Xiang, L., Fu, H., & Yang, C. (2025). CerfeVPR: Cross-environment robust feature enhancement for visual place recognition. Computers, Materials & Continua, 84(1), 325–345. https://doi.org/10.32604/cmc.2025.062834

[7] Zhao, X., Zhang, Y., Ning, J., Wang, Z., Li, K., & Zou, D. (2025). Semantic boundary constrained network for visual place recognition under adverse conditions. IEEE Transactions on Intelligent Transportation Systems, 26(8), 11823–11834. https://doi.org/10.1109/tits.2025.3576106

[8] Guan, P., Cao, Z., Fan, S., Yang, Y., Yu, J., & Wang, S. (2025). Hardness-aware metric learning with cluster-guided attention for visual place recognition. IEEE Transactions on Circuits and Systems for Video Technology, 35(1), 367–379. https://doi.org/10.1109/tcsvt.2024.3465520

[9] Shen, Y., Zhou, S., Fu, J., Wang, R., Chen, S., & Zheng, N. (2025). StructVPR++: Distill structural and semantic knowledge with weighting samples for visual place recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(8), 6338–6351. https://doi.org/10.1109/tpami.2025.3556859

[10] Ming, Y., Xu, M., Yang, X., Ye, W., Wang, W., Peng, Y., Dai, W., & Kong, W. (2025). VIPeR: Visual incremental place recognition with adaptive mining and continual learning. IEEE Robotics and Automation Letters, 10(3), 3038–3045. https://doi.org/10.1109/lra.2025.3539093

[11] Chen, C., He, M., Wang, J., & Meng, Z. (2024). GeoCluster: Enhancing visual place recognition in spatial domain on aerial vehicle platforms. IEEE Robotics and Automation Letters, 9(3), 3013–3020. https://doi.org/10.1109/lra.2024.3363536

[12] Grainge, O., Milford, M., Bodala, I., Ramchurn, S. D., & Ehsan, S. (2025). Structured pruning for efficient visual place recognition. IEEE Robotics and Automation Letters, 10(2), 2024–2031. https://doi.org/10.1109/lra.2024.3523222

[13] Xu, Y., Shamsolmoali, P., Zareapoor, M., & Yang, J. (2025). ClusVPR: Efficient visual place recognition with clustering-based weighted transformer. IEEE Transactions on Artificial Intelligence, 6(4), 1038–1049. https://doi.org/10.1109/tai.2024.3510479

[14] Jin, T., Lu, F., Hu, S., Yuan, C., & Liu, Y. (2025). EDTformer: An efficient decoder transformer for visual place recognition. IEEE Transactions on Circuits and Systems for Video Technology, 35(9), 8835–8848. https://doi.org/10.1109/tcsvt.2025.3559084

[15] Kannan, S. S., & Min, B.-C. (2024). PlaceFormer: Transformer-based visual place recognition using multi-scale patch selection and fusion. IEEE Robotics and Automation Letters, 9(7), 6552–6559. https://doi.org/10.1109/lra.2024.3408075

[16] Hou, S., Qin, Z., Wu, C., Yin, G., Wang, X., & Wang, Z. (2025). DIEONet: Domain-invariant information extraction and optimization network for visual place recognition. Computers, Materials & Continua, 82(3), 5019–5033. https://doi.org/10.32604/cmc.2025.058233

[17] Semenkov, I., Karpov, A., Savchenko, A. V., & Makarov, I. (2024). Inpainting semantic and depth features to improve visual place recognition in the wild. IEEE Access, 12, 5163–5176. https://doi.org/10.1109/access.2024.3350038

[18] Wang, S., Kang, Q., She, R., Zhao, K., Song, Y., & Tay, W. P. (2024). PRFusion: Toward effective and robust multi-modal place recognition with image and point cloud fusion. IEEE Transactions on Intelligent Transportation Systems, 25(12), 20523–20534. https://doi.org/10.1109/tits.2024.3465830

[19] Bao, Y., Pan, Y., Yang, Z., & Huan, R. (2024). A hierarchical utilization of semantic gradients and scene structure for visual place recognition. IEEE Transactions on Cognitive and Developmental Systems, 16(2), 570–583. https://doi.org/10.1109/tcds.2023.3281870

[20] Liu, L., Zuo, W., Fu, J., Shen, Y., Xin, J., & Zheng, N. (2025). RankTuning: Cross-image partial tuning strategies for rank optimization in visual place recognition. IEEE Transactions on Intelligent Transportation Systems, 26(12), 23179–23189. https://doi.org/10.1109/tits.2025.3614340

[21] Miller, E., Milford, M., Hafez, M. B., Ramchurn, S., & Ehsan, S. (2026). Through the lens of doubt: Robust and efficient uncertainty estimation for visual place recognition. IEEE Robotics and Automation Letters, 11(5), 5899–5906. https://doi.org/10.1109/lra.2026.3674688

[22] Berton, G., Trivigno, G., Caputo, B., & Masone, C. (2024). JIST: Joint image and sequence training for sequential visual place recognition. IEEE Robotics and Automation Letters, 9(2), 1310–1317. https://doi.org/10.1109/lra.2023.3339058

[23] Yuan, F., Schubert, S., Protzel, P., & Neubert, P. (2024). Local positional graphs and attentive local features for a data and runtime-efficient hierarchical place recognition pipeline. IEEE Robotics and Automation Letters, 9(3), 2686–2693. https://doi.org/10.1109/lra.2024.3359552

[24] Wenzel, P., Yang, N., Wang, R., Zeller, N., & Cremers, D. (2025). 4Seasons: Benchmarking visual SLAM and long-term localization for autonomous driving in challenging conditions. International Journal of Computer Vision, 133(4), 1564–1586. https://doi.org/10.1007/s11263-024-02230-4

[25] Li, P., Wen, S., Xu, C., & Qiu, T. Z. (2024). Visual place recognition for opposite viewpoints and environment changes. IEEE Transactions on Instrumentation and Measurement, 73, 1–9. https://doi.org/10.1109/tim.2024.3350152

[26] Zhao, J., Zhang, F., Cai, Y., Tian, G., Mu, W., Ye, C., & Feng, T. (2024). Learning sequence descriptor based on spatio-temporal attention for visual place recognition. IEEE Robotics and Automation Letters, 9(3), 2351–2358. https://doi.org/10.1109/lra.2024.3354627

[27] Lu, F., Dong, S., Zhang, L., Liu, B., Lan, X., Jiang, D., & Yuan, C. (2024). Deep homography estimation for visual place recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 38(9), 10341–10349. https://doi.org/10.1609/aaai.v38i9.28901

[28] Hausler, S., & Moghadam, P. (2025). Pair-VPR: Place-aware pre-training and contrastive pair classification for visual place recognition with vision transformers. IEEE Robotics and Automation Letters, 10(4), 4013–4020. https://doi.org/10.1109/lra.2025.3546512

[29] Yu, Z., Xu, Q., Jiang, Y., Sun, Y., & Huang, Q. (2024). Enhancing sample utilization in noise-robust deep metric learning with subgroup-based positive-pair selection. IEEE Transactions on Image Processing, 33, 6083–6097. https://doi.org/10.1109/tip.2024.3482182

[30] Li, H., Peng, G., Zhang, J., Wen, M., Ma, Y., & Wang, D. (2025). CaseVPR: Correlation-aware sequential embedding for sequence-to-frame visual place recognition. IEEE Robotics and Automation Letters, 10(4), 3430–3437. https://doi.org/10.1109/lra.2025.3541452

[31] Grainge, O., Milford, M., Bodala, I., Ramchurn, S. D., & Ehsan, S. (2024). Design space exploration of low-bit quantized neural networks for visual place recognition. IEEE Robotics and Automation Letters, 9(6), 5070–5077. https://doi.org/10.1109/lra.2024.3386459

[32] Hu, Z., Liao, L., & Lin, W. (2026). SRD2-VPR: Semantics-enforced feature aggregation with query rejection for visual place recognition. IEEE Transactions on Circuits and Systems for Video Technology, 36(5), 6934–6949. https://doi.org/10.1109/tcsvt.2026.3651681

[33] Soncini, N., Civera, J., & Pire, T. (2025). Addressing the challenges of loop detection in agricultural environments. Journal of Field Robotics, 42(1), 346–355. https://doi.org/10.1002/rob.22414

[34] Melekhin, A., Yudin, D. A., Petryashin, I., & Bezuglyj, V. (2025). MSSPlace: Multi-sensor place recognition with visual and text semantics. IEEE Access, 13, 177098–177110. https://doi.org/10.1109/access.2025.3618728