A Hyperdimensional One Place Signature to Represent Them All: Stackable Descriptors For Visual Place Recognition (Supplemental Material)

Connor Malone Somayeh Hussaini Tobias Fischer Michael Milford
QUT Centre for Robotics, Queensland University of Technology, Australia
{cj.malone, s.hussaini, tobias.fischer, michael.milford}@qut.edu.au

1. Overview

In the following document, we provide supplemental material containing extended results for the experiments presented in the main manuscript. We begin by providing recall@1 results using the AnyLoc [9] and BoQ [2] Visual Place Recognition (VPR) descriptors in Sections 2 and 3, followed in Section 4 by runtime evaluation to experimentally show the computational benefits of HOPS and prove it's equivalency to using a single reference set. Next, we provide experiments investigating the reduced benefits of HOPS for lower-dimensional descriptors in Section 5, and additional figures showing dimensionality reduction performance on more query sets in Section 6. Sections 7 and 8 provide results and discussion for the performance of HOPS on unstructured datasets (Google Landmarks and Pittsburgh 250k), with Section 9 expanding on results for the dataset identification experiment discussed in the main manuscript. We finish with additional experiments and discussion on other feature aggregation methods, the use of synthetic image augmentations with HOPS, and some qualitative VPR results in Sections 10, 11, and 12.

2. AnyLoc Recall@1

Core results in the main manuscript (Sections 4.3 and 4.4) demonstrate how our HOPS fused descriptors can improve the recall@1 performance of multiple state-of-the-art (SOTA) VPR descriptors across a range of adverse conditions. Here, we present the recall@1 results for another recent SOTA VPR descriptor, AnyLoc [9]. We again use the implementation provided in the VPR method evaluation repository released with EigenPlaces[1], based on the original implementation and using the author-released weights. We note that this implementation does not include the PCA component for dimensionality reduction presented by the authors, and we could not find released weights for PCA.

Table 1 shows VPR performance for AnyLoc using single reference sets, multi-reference set approaches, and our HOPS fused descriptor approach across all datasets used in the main manuscript. The tables demonstrate that the

Table 1. Recall@1 performance across all datasets for the AnyLoc VPR descriptor (49152D)

References ↓	Oxford RobotCar					
Queries →	Dusk	Night	Ov.cast	Ov.cast2	Rain	
Sunny	63.1	60.6	75.3	82.9	80.6	
Dusk	-	68.7	48.5	52.3	57.4	
Night	69.7	-	49.4	49.4	51.7	
Overcast	67.5	64.9	-	83.0	81.1	
Overcast2	65.6	61.7	79.2	-	80.9	
Rain	66.9	58.7	74.0	78.0	-	
dMat Avg [7]	81.0	77.1	80.6	86.3	85.8	
Pooling	73.9	68.7	83.1	87.5	87.1	
HOPS (Ours)	**84.0**	**79.0**	**86.9**	**90.7**	**91.4**	
	Nordland					
Queries →	Fall	Spring	Sum.	Winter		
Fall	-	62.0	70.5	37.1		
Spring	59.9	-	56.8	32.7		
Summer	70.9	57.1	-	33.0		
Winter	24.7	36.2	22.4	-		
dMat Avg [7]	73.5	71.1	71.4	47.9		
Pooling	73.4	66.9	71.2	35.5		
HOPS (Ours)	**78.0**	**76.5**	**75.1**	**48.1**		
	SFU-Mountain					
Queries →	Dry	Dusk	Jan	Nov	Sept	Wet
Dry	-	77.7	57.4	66.0	62.1	68.1
Dusk	84.7	-	72.5	73.5	66.8	89.4
Jan	56.1	69.6	-	60.0	52.7	70.1
Nov	68.3	62.9	62.3	-	71.7	65.7
Sept	61.3	59.7	56.6	71.7	-	62.6
Wet	77.7	91.2	70.9	77.4	63.9	-
dMat Avg [7]	91.4	95.8	92.5	92.2	87.3	93.5
Pooling	88.1	93.2	82.6	82.9	79.2	91.2
HOPS (Ours)	**97.4**	**98.4**	**93.5**	**97.1**	**92.5**	**97.1**

[1] https://github.com/gmberton/VPR-methods-evaluation

improvements to results observed for other SOTA VPR descriptors hold for AnyLoc as well, improving the recall@1 over the best single reference set recalls by at least absolute 7.7%, 4.6%, and 7.2%, and up to 14.3%, 14.6%, and 21%, respectively for the Oxford RobotCar [10], Nordland [12], and SFU Mountain datasets [6].

For these AnyLoc results, our HOPS fused descriptors achieve higher recall@1 than **both** the best single reference set **and** the other multi-reference set approaches in **all** cases. AnyLoc could be particularly suited to the use of hyperdimensional computing frameworks due to the large dimensionality of its feature vectors (49152D) prior to any dimensionality reduction.

3. BoQ Recall@1

Following on from the above AnyLoc results, we also provide results for HOPS with BoQ [2] VPR descriptors (using the DinoV2 model).

Table 2 shows a near unanimous improvement in Recall@1 over both the best single reference set results and the alternative multi-reference set approaches. The SFU Mountain 'Wet' query set is the only condition where HOPS is not the best performing method. Similarly to SALAD, CricaVPR, and AnyLoc, the high dimensionality of BoQ features appears to be well suited to our HOPS fused descriptors.

4. Runtime Evaluation

Using the theoretical computational complexity, it can be asserted that our HOPS descriptors do not increase the computational overhead of performing VPR. Accordingly, HOPS provides significant advantage over other multi-reference VPR approaches such as distance matrix averaging [7] or reference set pooling, which both increase computational overheads with a complexity of $\mathcal{O}(K \cdot M)$. We empirically verified this claim using the SALAD VPR descriptor on the RobotCar datasets. All fusion methods shared a query feature extraction time of 11.1ms. The distance matrix averaging and reference set pooling approaches had image matching times of 53.7ms and 52.3ms respectively, whereas, the baseline (single reference) approach and our HOPS descriptors both had a image matching time of 10.6ms. These results are intended to provide evidence of the relative runtime difference between approaches, however for completeness, they were evaluated on an Ubuntu desktop using an Intel i7-12700K CPU, 32GB of RAM, and an NVIDIA GeForce RTX 3080 Ti GPU.

5. Lower Dimensional Descriptors

In the main manuscript, it can be observed that our HOPS fused descriptors do not provide an advantage as consistently for lower dimensional descriptors (512D), such as CosPlace [4] and EigenPlaces [5], compared to higher dimensional ones. To investigate these results further, we evaluate the performance of CosPlace on the RobotCar dataset using the extended range of descriptor sizes available[2].

Table 2. Recall@1 performance across all datasets for the BoQ VPR descriptor (12288D)

References ↓	**Oxford RobotCar**					
Queries →	**Dusk**	**Night**	**Over.**	**Over.2**	**Rain**	
Sunny	80.2	77.9	87.9	90.5	89.1	
Dusk	-	76.8	77.0	76.6	78.7	
Night	76.0	-	73.9	71.8	71.7	
Overcast	81.6	79.6	-	90.4	90.2	
Overcast2	82.1	75.7	90.0	-	89.7	
Rain	81.8	76.2	89.0	89.1	-	
dMat Avg	89.6	86.7	92.6	93.5	93.1	
Pooling	85.8	80.5	92.2	93.1	92.8	
HOPS (Ours)	**90.7**	**87.4**	**93.9**	**94.4**	**94.1**	
	Nordland					
Queries →	**Fall**	**Spring**	**Summer**	**Winter**		
Fall	-	80.8	79.8	76.2		
Spring	79.5	-	78.2	79.1		
Summer	80.0	78.7	-	73.7		
Winter	73.4	79.5	72.5	-		
dMat Avg	81.8	82.3	80.7	80.9		
Pooling	81.7	82.3	80.8	80.0		
HOPS (Ours)	**82.4**	**82.6**	**81.0**	**81.2**		
	SFU-Mountain					
Queries →	**Dry**	**Dusk**	**Jan**	**Nov**	**Sept**	**Wet**
Dry	-	100.0	96.4	97.1	96.1	97.9
Dusk	99.5	-	97.4	96.6	96.4	99.2
Jan	96.6	98.2	-	96.6	95.3	96.1
Nov	98.4	97.1	96.6	-	97.4	95.6
Sept	94.3	92.7	93.5	96.9	-	92.7
Wet	98.7	99.5	97.1	96.9	96.9	-
dMat Avg	99.5	**100.0**	98.2	98.7	97.7	99.0
Pooling	99.5	**100.0**	98.2	98.7	97.7	**99.5**
HOPS (Ours)	**99.7**	**100.0**	**98.7**	**99.5**	**98.2**	99.0

Table 3. Recall@1 on RobotCar using CosPlace

Queries →	**Dusk**	**Night**	**Ovr**	**Ovr2**	**Rain**
Best Single Ref (64D)	33.2	10.9	71.5	78.0	73.3
HOPS (Ours) (64D)	40.2	7.3	72.3	78.3	76.3
Best Single Ref (512D)	48.6	21.2	84.2	86.5	84.6
HOPS (Ours) (512D)	57.0	19.4	85.9	89.8	90.2
Best Single Ref (2048D)	50.7	19.8	83.0	87.2	85.0
HOPS (Ours) (2048D)	57.7	16.3	86.6	90.2	89.9

[2] https://github.com/gmberton/VPR-methods-evaluation

Table 3 shows that CosPlace achieves a consistent improvement for all RobotCar query sets across all dimensionalities except night-time, where a consistent *decrease* in performance is seen. This indicates that the classification-style training used for CosPlace or the night-time conditions may be the cause of reduced performance noted in the main manuscript, rather than lower dimensionality.

6. Dimensionality Reduction

In the main manuscript, we investigated how recall@1 performance is affected by reducing the dimensionality of descriptors using a Gaussian Random Projection and presented results for the Oxford RobotCar Dusk query set. Here, we extend these results and provide the corresponding figures for all six VPR descriptors across two additional query sets from each dataset (Figures 1–6). The additional results confirm the finding of the main manuscript and demonstrate that our HOPS fused descriptors can achieve the same or better performance compared to using the best single reference set, with up to a 97% reduction in descriptor dimensionality. Despite the low dimensionality of the CosPlace [4] and EigenPlaces [5] descriptors (512D), our HOPS fused descriptors still match or exceed the best single reference recall with reduced dimensionality for 9/12 cases shown.

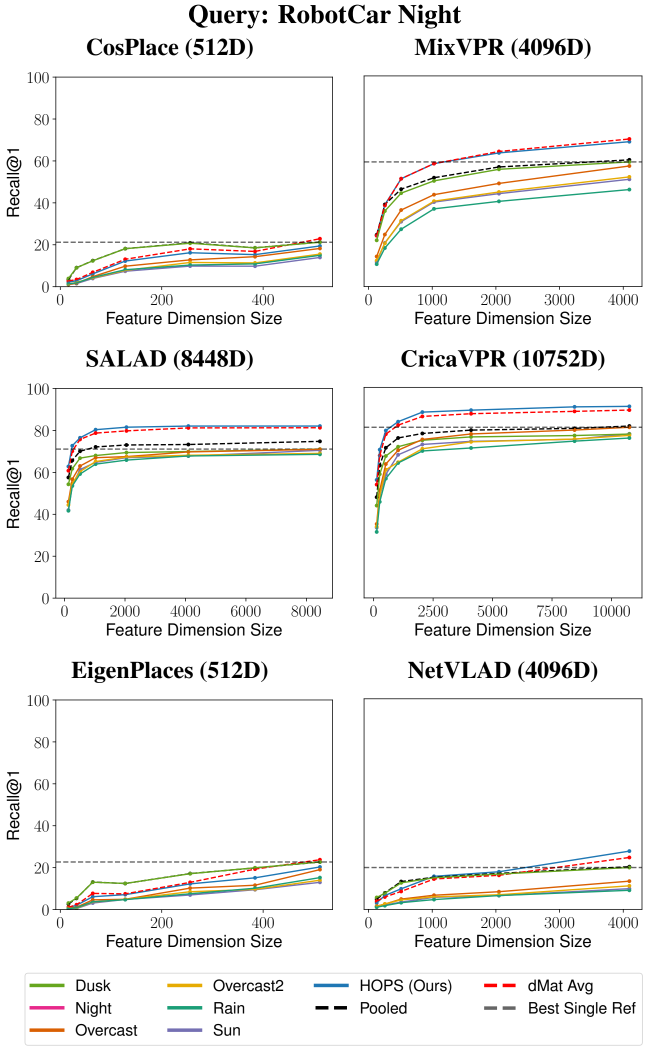

Figure 1. Recall@1 performance for different VPR descriptors across the Oxford RobotCar Night set as dimensionality is reduced using Gaussian Random Projection.

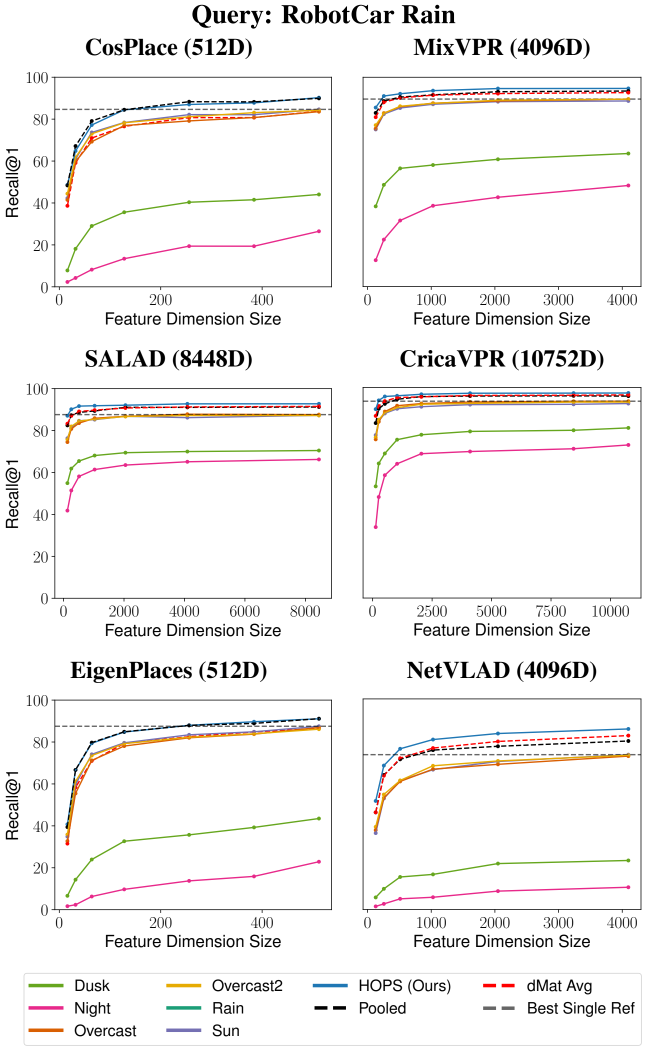

Figure 2. Recall@1 performance for different VPR descriptors across the Oxford RobotCar Rain set as dimensionality is reduced using Gaussian Random Projection.

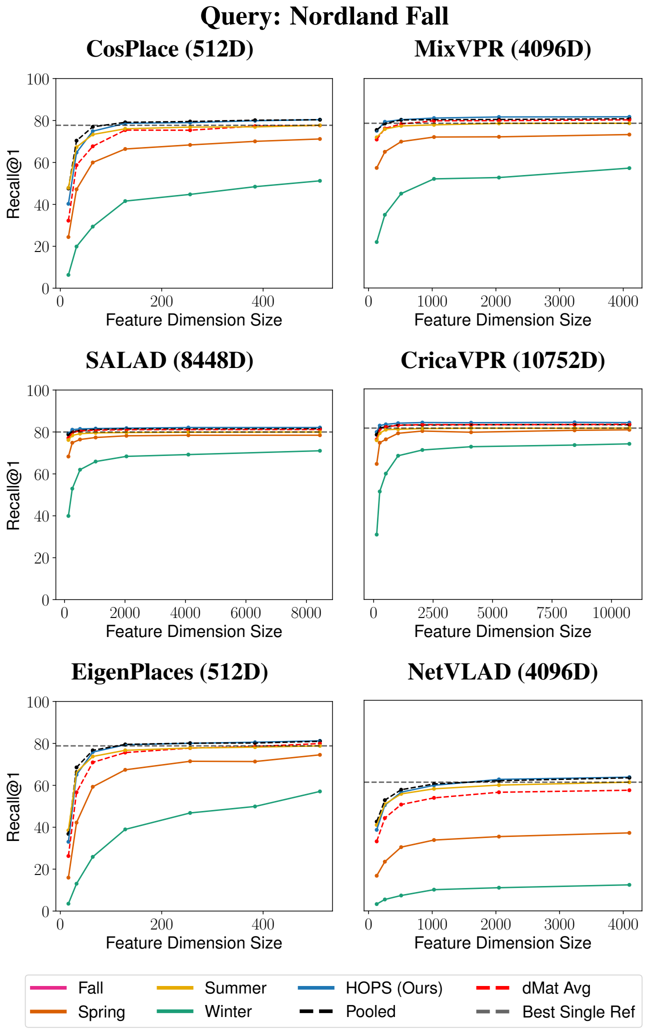

Figure 3. Recall@1 performance for different VPR descriptors across the Nordland Fall set as dimensionality is reduced using Gaussian Random Projection.

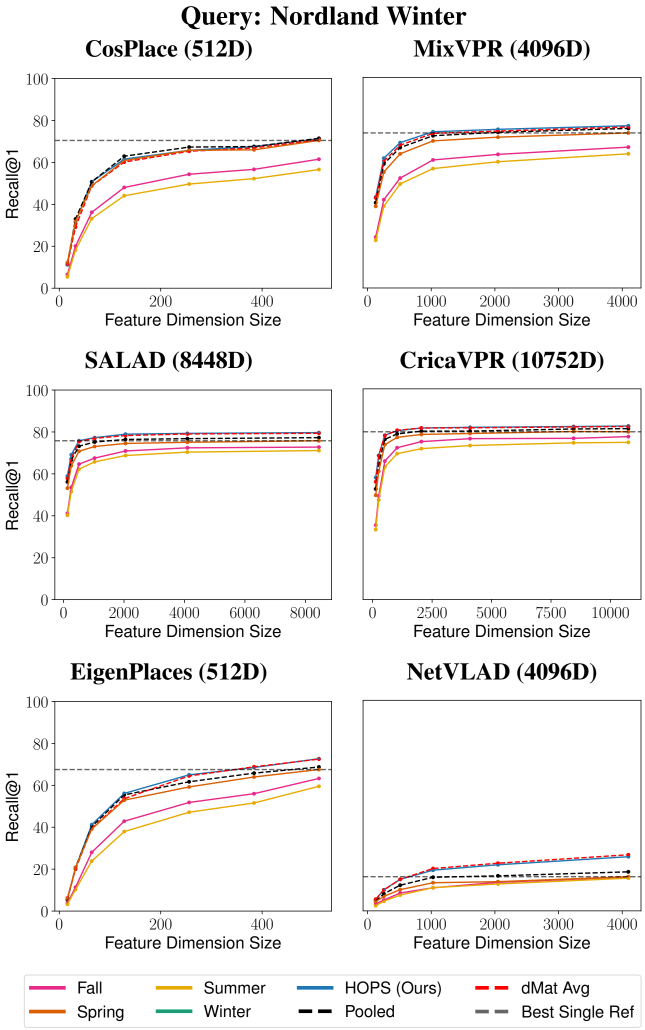

Figure 4. Recall@1 performance for different VPR descriptors across the Nordland Winter set as dimensionality is reduced using Gaussian Random Projection.

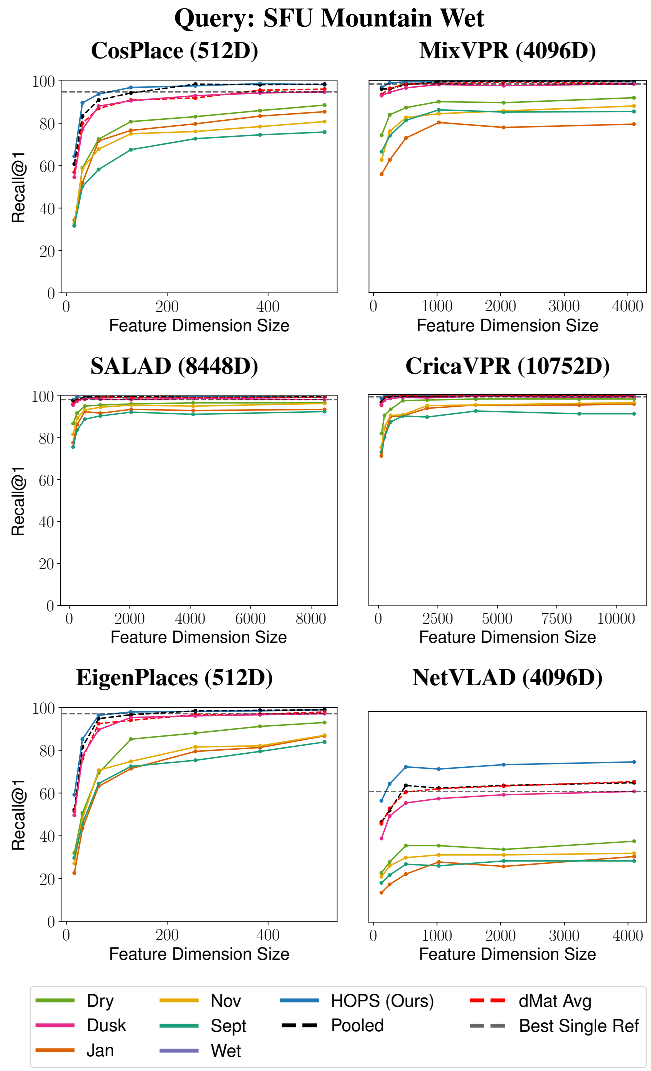

Figure 5. Recall@1 performance for different VPR descriptors across the SFU Mountain Wet set as dimensionality is reduced using Gaussian Random Projection.

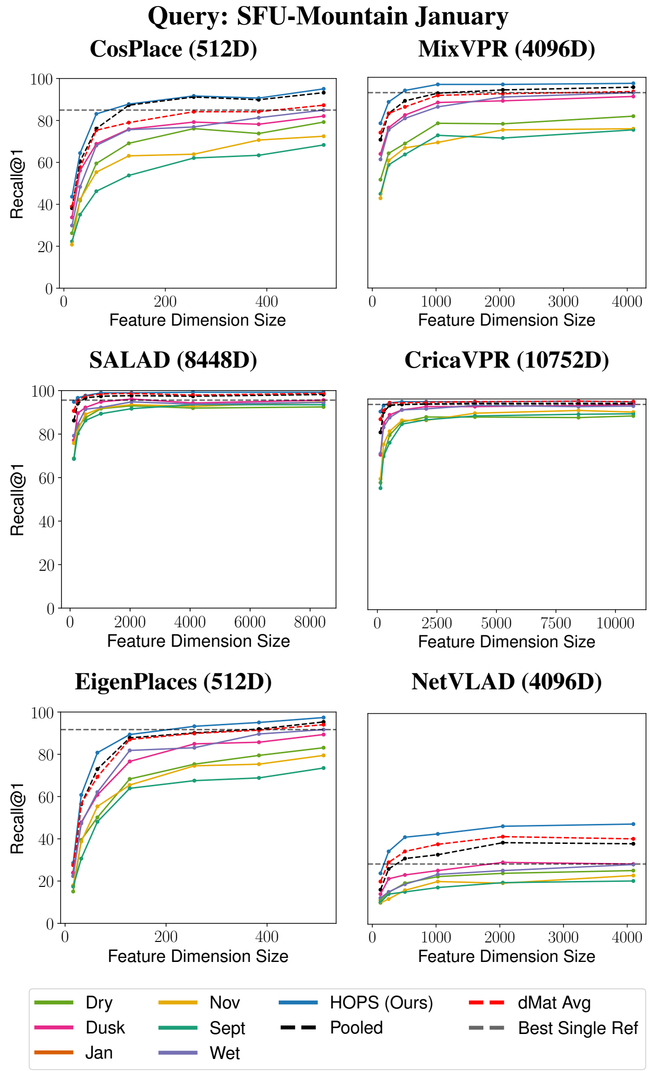

Figure 6. Recall@1 performance for different VPR descriptors across the SFU Mountain January set as dimensionality is reduced using Gaussian Random Projection.

7. Unstructured Datasets: Google Landmarks

One of the major challenges in robotics, and VPR, is maintaining performance in dynamic and unstructured environments. Datasets and results included in the main manuscript already contain extensive instances of dynamic environments, with environmental condition changes evident in all datasets (i.e. Nordland, Oxford RobotCar, SFU Mountain). For example, the Oxford RobotCar [10] dataset contains weather/time of day changes, dynamic objects (i.e. cars and pedestrians), and temporal changes such as construction and infrastructure changes. All of these datasets contain images captured from the same route throughout an environment which could be considered independently as either query or reference sets. However, some datasets are unstructured and contain a collection of non-sequential images, with a varied number of images per 'place', which can be localized using meta data such as latitude, longitude, and heading.

In this section, we present and discuss results for using our HOPS fused descriptors to compress the more unstructured reference dataset for the Google Landmarks v2 micro dataset. This version of the Google Landmarks v2 dataset contains $23,294$ reference images and $3,103$ query images, each labeled with a 'Landmark ID'. Using our HOPS approach, we were able to fuse 7-9 reference images with the same 'Landmark ID' from each place and reduce the reference set to just $3,103$ images; 13.3% of the original size. Using the SALAD VPR descriptor [8], Table 4 shows that HOPS fused descriptors are able to substantially reduce the reference set size while only incurring a small decrease in Recall@1 performance (3.9%), therefore *significantly* reducing compute and storage requirements. These results were reflected in further evaluation using BoQ, CricaVPR, and MixVPR descriptors, where the same dataset compression could be achieved with similarly small reductions in Recall@1 (2.3%, 4.9%, and 5.1% respectively).

For comparison, we also provide results for an alternative strategy where dimensionality reduction is used to reduce the reference set size rather than our HOPS descriptors. We reduce feature vectors to 1024 dimensional to provide equivalent memory requirements compared to our HOPS reduced reference set. Table 4 shows that our HOPS fused descriptors are able to maintain a much higher Recall@1 at this memory footprint compared to using dimensionality reduction methods.

Table 4. Recall@1 performance using SALAD [8] on the Google Landmarks v2 micro dataset for different reference set reduction strategies. Our HOPS fused descriptors significantly reduce reference set size while only incurring a small decrease in Recall@1.

Reference Set	Original	HOPS Reduced	Dim-Reduced Feats.
Num. of Refs.	23,294	3,103	23,294
Feat. Dim.	8488	8448	1024
Recall@1	69.7	65.8	59.7

8. Unstructured Datasets: Pittsburgh 250k

In addition to results for the compression of the Google Landmarks dataset, we also evaluate the performance of HOPS descriptors in a similar experiment using the Pittsburgh 250k dataset [13]. The Pittsburgh dataset consists of $\approx 250,000$ images which provide dense, large-scale coverage of the city and capture multiple different viewpoints at each singular location. In this section, we evaluate the performance of our HOPS descriptors for fusing reference images that are spatially close to each other, do not vary in appearance condition, but may slightly vary in viewpoint. This differs from the previous experiment using the Google Landmarks dataset, where fused images captured the same 'landmark' or features but from differing viewpoints.

For this experiment, we use the test subset of the Pittsburgh dataset with 83,952 reference images and 24,000 queries from 1,000 unique locations, each captured from 24 different viewpoints. To generate our HOPS descriptors, we fuse descriptors from reference images that are within a 25m distance radius and have a cosine similarity higher than 0.5. We use cosine similarity as a proxy to filter images that are spatially close but not captured from a similar viewpoint (e.g. viewing the opposite direction down a street). This fusion replaces the reference descriptors that have highly visually similar neighbors nearby with their HOPS descriptors. Each HOPS descriptor is obtained by fusing a reference descriptor with neighboring descriptors that meet the outlined criteria. As a result, the total number of reference descriptors remains the same, and neighboring descriptors are potentially fused into multiple different HOPS descriptors. A total of 21,831 reference descriptors are replaced, while the remaining 62,121 reference descriptors remain unchanged.

Using MixVPR, HOPS descriptors only incur an absolute decrease of 0.38% in Recall@1 (from 94.28% to 93.9%). This suggests that small viewpoint differences between neighboring images can introduce slight noise into the fused HOPS descriptor. This also indicates that *increases* in Recall@1 using HOPS may only be experienced when using multi-condition reference sets to improve robustness in adverse conditions (as demonstrated in the main manuscript). There is a significant opportunity for further exploration of how and when HOPS can be used to improve VPR performance. However, the lack of datasets that are dense, contain repeated locations captured under varied environmental conditions, *and* contain repeated locations from different view points, currently make a thorough investigation difficult.

9. Dataset Identification

Beyond fusing descriptors from the same place, there are many other possible applications for the HDC framework in VPR, such as dataset/environment identification. This

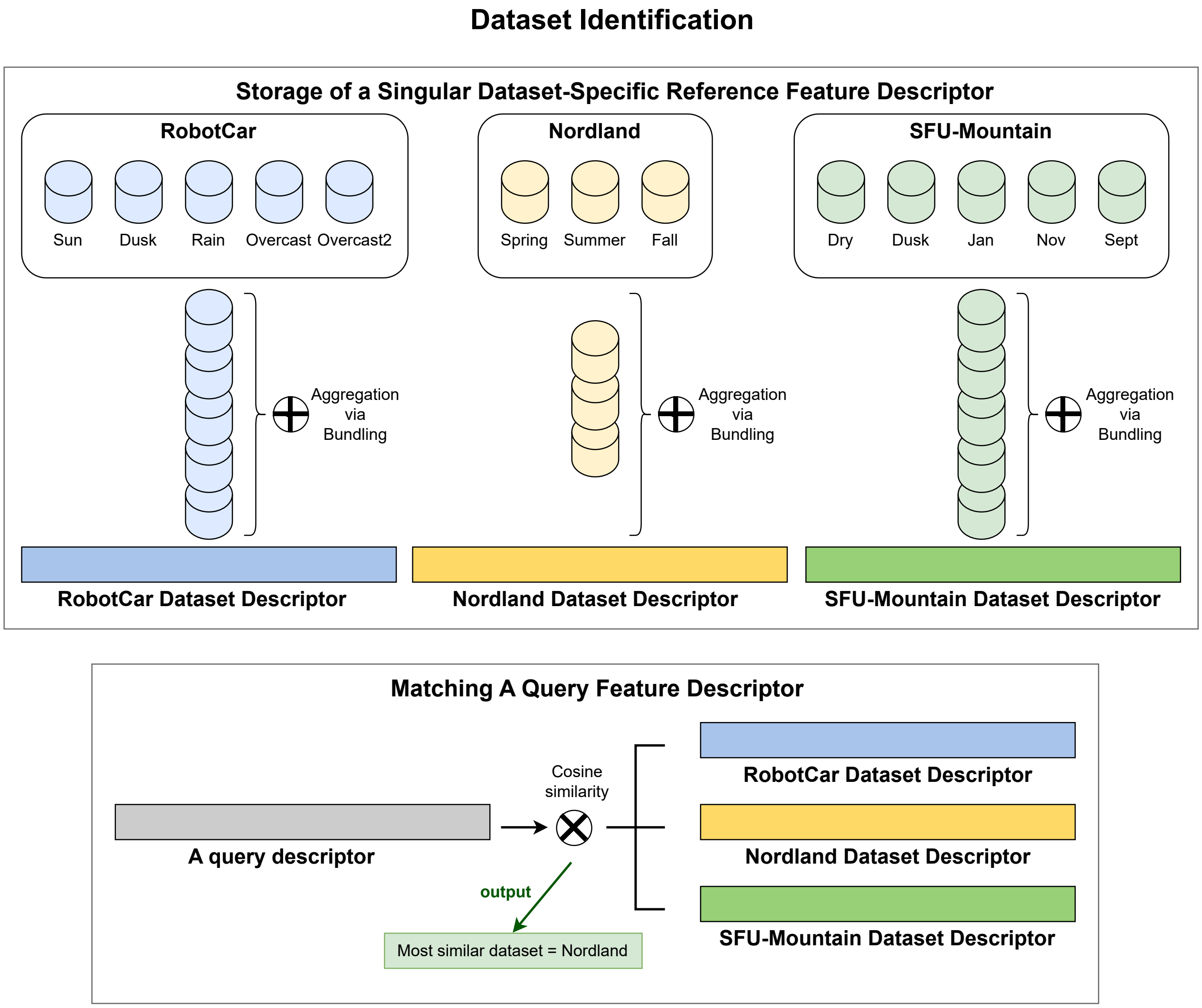
Dataset Identification
Storage of a Singular Dataset-Specific Reference Feature Descriptor
RobotCar
Sun
Dusk
Rain
Overcast
Overcast2
Nordland
Spring
Summer
Fall
SFU-Mountain
Dry
Dusk
Jan
Nov
Sept
Aggregation via Bundling
Aggregation via Bundling
Aggregation via Bundling
RobotCar Dataset Descriptor
Nordland Dataset Descriptor
SFU-Mountain Dataset Descriptor
Matching A Query Feature Descriptor
Cosine similarity
A query descriptor
output
Most similar dataset = Nordland
RobotCar Dataset Descriptor
Nordland Dataset Descriptor
SFU-Mountain Dataset Descriptor

Figure 7. The visualisation of our dataset identification investigation, as discussed earlier in Section 4.7. **Top:** Our HOPS fused descriptors aggregate the reference descriptors from each resepective dataset into a single descriptor to represent the entire dataset, essentially summarizing each dataset via a single representation. **Bottom:** To classify a query descriptor, the cosine similarity to each (fused) reference dataset descriptor is computed, with the highest cosine similarity indicating the predicted datasetfor the query.

section provides additional details on using our HOPS fused descriptors to identify which dataset a given query descriptor belongs to, as discussed in Section 4.7 of the main manuscript. Figure 7 provides a visualization of the overall process we used to perform this experiment.

First, for each dataset, we pool all respective available reference sets together. Then, we use the HDC bundling operation to aggregate *all* descriptors into a *single overall dataset descriptor*. After separately performing the bundling for the three datasets, RobotCar, Nordland, and SFU Mountain, this provides three ‘dataset’ reference descriptors.

To evaluate the performance of our *dataset-specific fused descriptor*, we identified the source dataset of each query descriptor by calculating its cosine similarity against each of the single overall dataset descriptors. We emphasize that when evaluating the accuracy of dataset identification for each query set, the respective set was removed from the bundling operation and therefore was not included in the overall dataset descriptors.

To provide context for these results, we compare against alternate dataset identification approaches. The first method we compare to is one where a *single* descriptor is randomly chosen to represent each dataset from the respective reference sets. We also evaluate dataset identification accuracy when 10 descriptors are randomly chosen per dataset, covering multiple reference sets. For this approach, the query

Table 5. Dataset identification accuracy across all datasets using the SALAD VPR descriptor.

	Oxford RobotCar					Nordland				SFU-Mountain					
Queries →	**Dusk**	**Night**	**Overcast**	**Overcast2**	**Rain**	**Fall**	**Spring**	**Summer**	**Winter**	**Dry**	**Dusk**	**Jan**	**Nov**	**Sept**	**Wet**
Single Image	95.54	62.41	63.67	84.49	76.37	79.40	94.89	8.91	64.30	85.97	97.92	98.96	100	96.10	97.14
Pooled (10 Images)	94.61	96.88	93.19	95.82	97.65	95.19	96.96	94.72	94.84	100	100	100	100	100	100
HOPS (Ours)	100	100	99.85	99.79	99.79	99.97	99.82	99.87	99.8	100	100	99.74	100	100	100

descriptor is compared to all 10 reference descriptors from each dataset and the predicted dataset becomes the one which contains the reference descriptor most similar to the query according to cosine similarity.

Table 5 provides the accuracy of our HOPS fused descriptors and all comparisons for the dataset identification task using the SALAD VPR descriptor. It demonstrates that our approach can correctly predict the source dataset of a query image, with an accuracy of above 99.7% across all datasets, which is significantly better than a random single image and 10 random images per dataset (20.5% and 2.6% improvement on average respectively).

This experiment shows that our HOPS fused descriptor is able to distinguish between different environments based on their overall feature descriptor characteristics. However, these three datasets could be considered to be relatively distinct from each other and therefore easily identifiable. To further test the capability of HOPS descriptors for dataset identification, we evaluate accuracy using the Pittsburgh 30k [13] and Tokyo Time Machine [3] datasets. Both datasets are subsets of Google StreetView and therefore captured from a much more similar distribution. When used to determine if queries belong to either the Pittsburgh or Tokyo datasets, our HOPS fused descriptors achieve an accuracy of 98.2%. This is further evidence of HOPS' utility for dataset identification.

10. Other Feature Aggregation Methods

In the main manuscript, we compare our HOPS fused descriptors to the distance matrix averaging feature aggregation approach because it was the highest performing method from [7]. In this section, we provide a comparison to the other feature aggregation methods explored in [7]. These include taking the *minimum* values, *maximum* values, or *median* values from the distance matrix rather than the mean/average. We present results obtained using MixVPR [1] VPR descriptors on the Oxford RobotCar datasets [10].

Similarly to results seen in [7], Table 6 shows that the distance matrix averaging generally achieves equivalent or higher Recall@1 compared to the other feature aggregation methods from [7]. Notably, the 'Minimum' method achieves similar results on the 'Overcast' and 'Rain' datasets but considerably lower Recall@1 for more challenging conditions such as 'Dusk' and 'Rain'. Importantly, our HOPS fused descriptors generally maintain the highest Recall@1 out of all methods.

Table 6. Recall@1 on the RobotCar datasets using MixVPR and different feature aggregation methods. In addition to methods compared in the main manuscript, we present other approaches also explored in [7]. The results show the dMat averaging aggregation generally achieves higher Recall@1 across the different datasets compared to other methods from [7]; with HOPS nearly always achieving the highest Recall@1 of all methods.

Ref	**Sun**	**Dusk**	**Night**	**O/C**	**O/C2**	**Rain**	**Avg [22]**	**Pool**	**HOPS**	**Min [22]**	**Max [22]**	**Med [22]**
Qry ↓					**MixVPR (4096D)**							
Dusk	69.0	-	64.6	71.7	67.4	68.3	82.9	77.1	**83.1**	77.1	65.2	76.7
Night	50.9	59.2	-	57.2	52.0	46.0	**70.0**	60.1	68.8	60.1	53.8	60.3
O/C	86.3	60.1	52.2	-	89.1	87.1	91.5	92.0	**93.3**	92.0	52.6	85.9
O/C2	91.2	61.4	50.3	90.6	-	88.8	93.6	93.7	**94.7**	93.7	51.6	88.8
Rain	88.7	63.6	48.3	89.6	89.5	-	92.7	93.4	**94.7**	93.4	50.8	88.1

We would also like to reiterate the key advantages of HOPS compared to these other feature aggregation methods. HOPS fuses place descriptors while all methods from [7] fuse difference matrices obtained by running VPR on K reference-query image pairs per place; requiring to store K descriptors per place. [7]'s vanilla version and [11] need to compute K query image descriptors at inference, giving HOPS both significant computation (single query descriptor) and memory (single reference vector) advantages, and higher performance.

11. Synthetic Image Augmentations

In the main manuscript, we presented results for a proof-of-concept study where differing conditions can be substituted for synthetic augmentations of a reference set to create HOPS descriptors. This would enable more robust performance across environmental conditions without the need to collect multiple *real* reference sets. Results using SALAD VPR descriptors on the RobotCar datasets showed that this generally resulted in minor improvements to Recall@1, with the exception of a slight decrease being seen for the Overcast query set. In this section, we provide results for the

Table 7. Recall@1 on RobotCar datasets Using Synthetic Changes

Queries → **References ↓**	**Dusk**	**Night**	**Overcast**	**Overcast2**	**Rain**
			CosPlace (512D)		
Sun	44.1	14.0	78.3	86.5	84.6
Synthetic Dark	40.5	15.1	57.3	66.0	64.7
Poisson Noise	37.3	16.8	54.1	61.6	60.0
Downsample-Upsample	18.9	7.6	41.8	51.3	47.3
dMat Avg	41.8	16.6	68.3	78.0	74.8
Pooling	**43.6**	**17.1**	**78.3**	**86.5**	**84.6**
HOPS (Ours)	40.0	17.0	63.2	74.3	71.0
			MixVPR (4096D)		
Sun	70.1	52.4	86.5	91.0	88.5
Synthetic Dark	64.3	42.8	69.0	75.5	75.9
Poisson Noise	68.4	52.5	84.7	90.4	87.2
Downsample-Upsample	67.4	50.0	83.4	89.7	87.3
dMat Avg	**71.3**	**53.5**	**84.0**	89.5	86.9
Pooling	63.1	45.9	82.6	87.4	78.7
HOPS (Ours)	69.7	51.2	**84.0**	**90.1**	**88.0**
			CricaVPR (10752D)		
Sun	81.4	77.9	90.6	93.9	92.4
Synthetic Dark	67.0	60.1	68.9	76.1	73.4
Poisson Noise	70.1	66.2	82.1	88.5	85.1
Downsample-Upsample	69.5	66.4	81.5	88.4	84.7
dMat Avg	**76.2**	**71.4**	**84.3**	**90.1**	**87.3**
Pooling	68.9	64.1	80.7	88.4	80.2
HOPS (Ours)	71.9	67.1	81.8	88.6	84.2

CosPlace, MixVPR, and CricaVPR descriptors to further investigate the capability of synthetic augmentations with HOPS.

Table 7 shows that, in contrast to the minor improvements observed for SALAD descriptors, synthetic augmentations result in performance decreases for CosPlace, MixVPR, and CricaVPR. When performance was particularly poor, such as CosPlace on the night-time query set, HOPS using synthetic augmentations still made improvements, however, performance generally decreased. This may indicate that the synthetic augmentations are not generalizable across VPR descriptors. Future work could investigate the use of more 'realistic' augmentations produced by methods such as diffusion to determine if improvements could be made across a larger range of VPR descriptors and environmental conditions.

12. Qualitative Results

In this section, we provide qualitative results to show scenarios where our method excels and instances where it fails, evaluated across a range of diverse and challenging conditions. Figures 8, 9, and 10 show qualitative results on the RobotCar, Nordland and SFU Mountain datasets, respectively. We use green borders to indicate correct matches, and red borders to indicate false matches. In these figures, we show cases where our HOPS fused descriptors are able to retrieve correct matches even in cases where all other methods fail. We also include cases where HOPS fails to retrieve a correct match, while other methods either retrieve correct matches or also fail.

For comparability, we visualize VPR matches for all approaches using corresponding images from the single reference set which achieves the best recall@1; noting that for multi-reference methods such as the distance matrix averaging and our HOPS fused descriptors, the matches rely on multiple reference sets.

We reiterate that we use a tight ground truth tolerance of ± 2 frames for the RobotCar dataset, 0 frames for the Nordland dataset, and ± 1 frames for the SFU Mountain dataset. Therefore, while VPR methods are generally good at finding matches close to the ground truth location, our HOPS fused descriptors are able to further reduce the match errors, even those near the true match, disambiguating spatially close places. For example, this improvement is evident in the RobotCar dataset in Figure 8 rows 2 and 5, the Nordland dataset in Figure 9 rows 4 and 6, and the SFU Mountain dataset in Figure 10 rows 3 and 6.

We have also included example visualizations where all or the majority of methods fail to match to the correct places, which are mostly due to high visual similarity between geographically distant places. These instances are shown for RobotCar, Nordland and SFU Mountain datasets in rows 7 and 8 of Figures 8, 9, and 10, respectively.

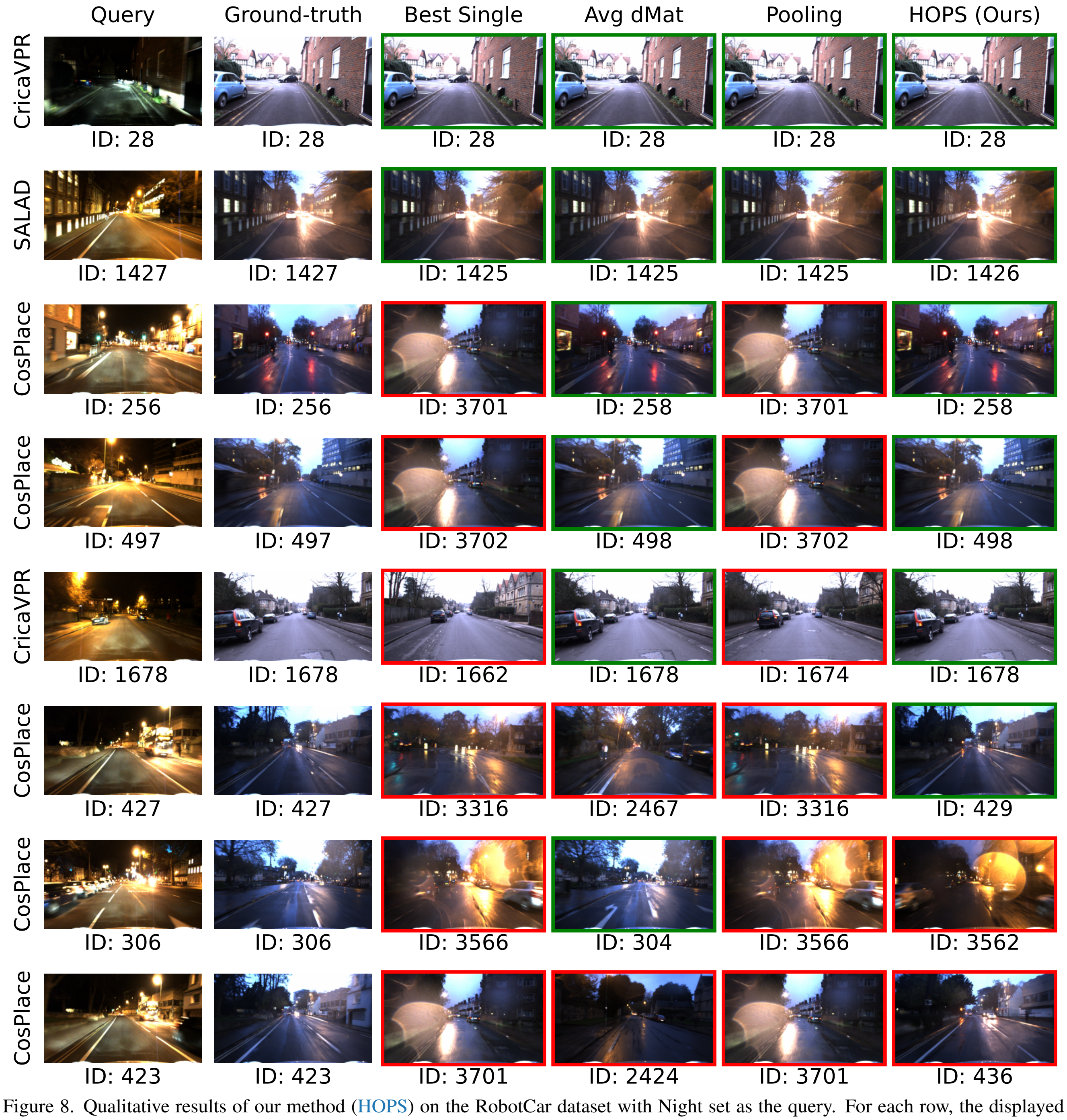

Figure 8. Qualitative results of our method (HOPS) on the RobotCar dataset with Night set as the query. For each row, the displayed reference set corresponds to the 'Best Single' reference set for the specific method, shown to the right of the query image. We show three different scenarios: i. Cases where fusing reference sets via HOPS (ours) produces correct matches, similar to other techniques (rows 1 and 2). ii. Cases where HOPS produces correct matches while at least one other method fails (rows 3, 4, 5, and 6). iii. Cases where HOPS retrieves false matches and other methods either succeed or fail (rows 7 and 8). Note that we use a tight ground truth tolerance of ± 2 meters for the RobotCar dataset. HOPS fused descriptors further reduce the error of matches already made in close proximity to the true match, effectively disambiguating spatially close places.

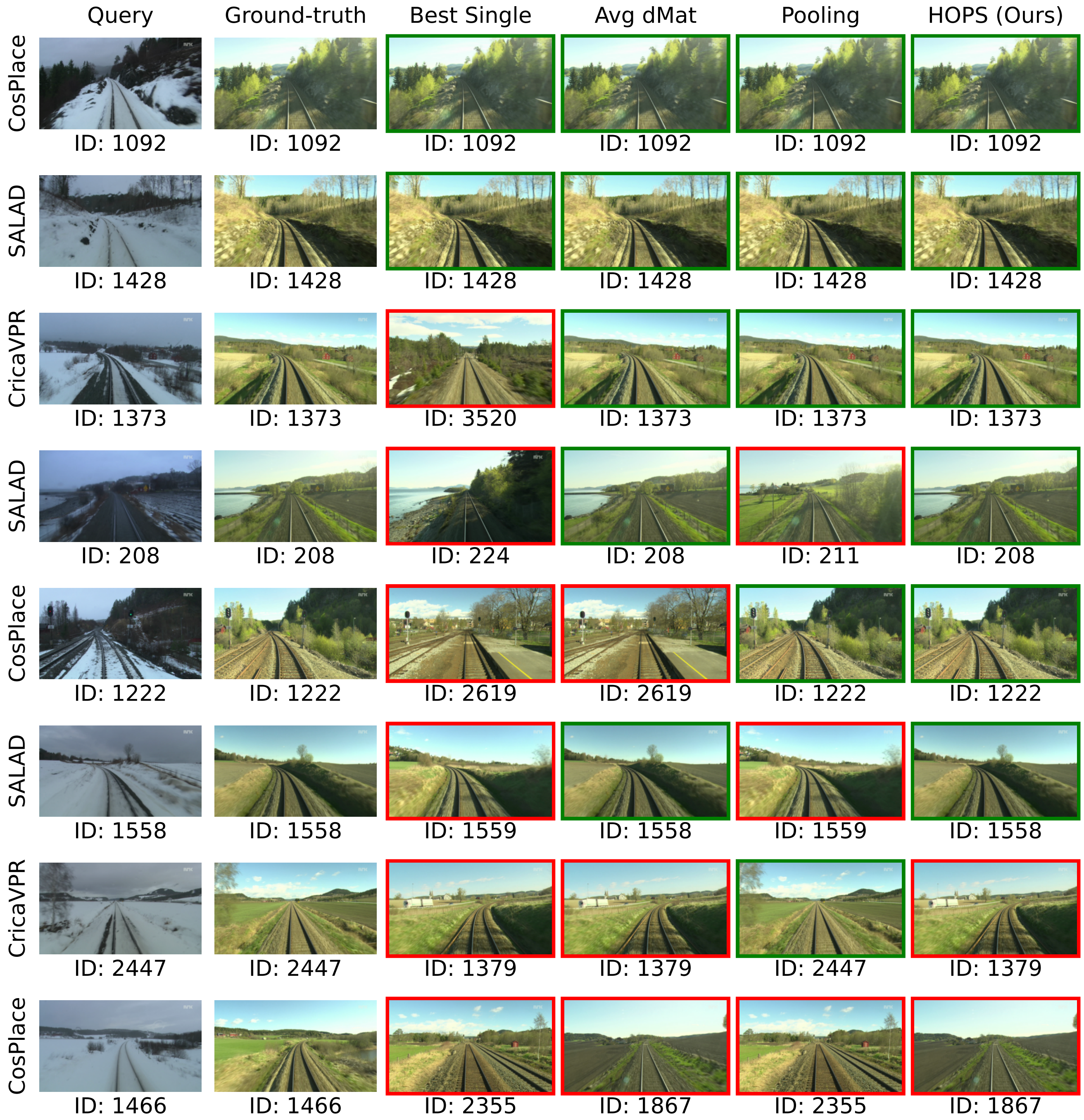

Figure 9. Qualitative results of our method (HOPS) on the Nordland dataset with Winter set as the query. For each row, the displayed reference set corresponds to the 'Best Single' reference set for the specific method, shown to the right of the query image. We show three different scenarios: i. Cases where fusing reference sets via HOPS (ours) produces correct matches, similar to other techniques (rows 1 and 2). ii. Cases where HOPS produces correct matches while at least one other method fails (rows 3, 4, 5, and 6). iii. Cases where HOPS retrieves false matches and other methods either succeed or fail (rows 7 and 8). Note that we use a tight ground truth tolerance of 0 images for the Nordland dataset.

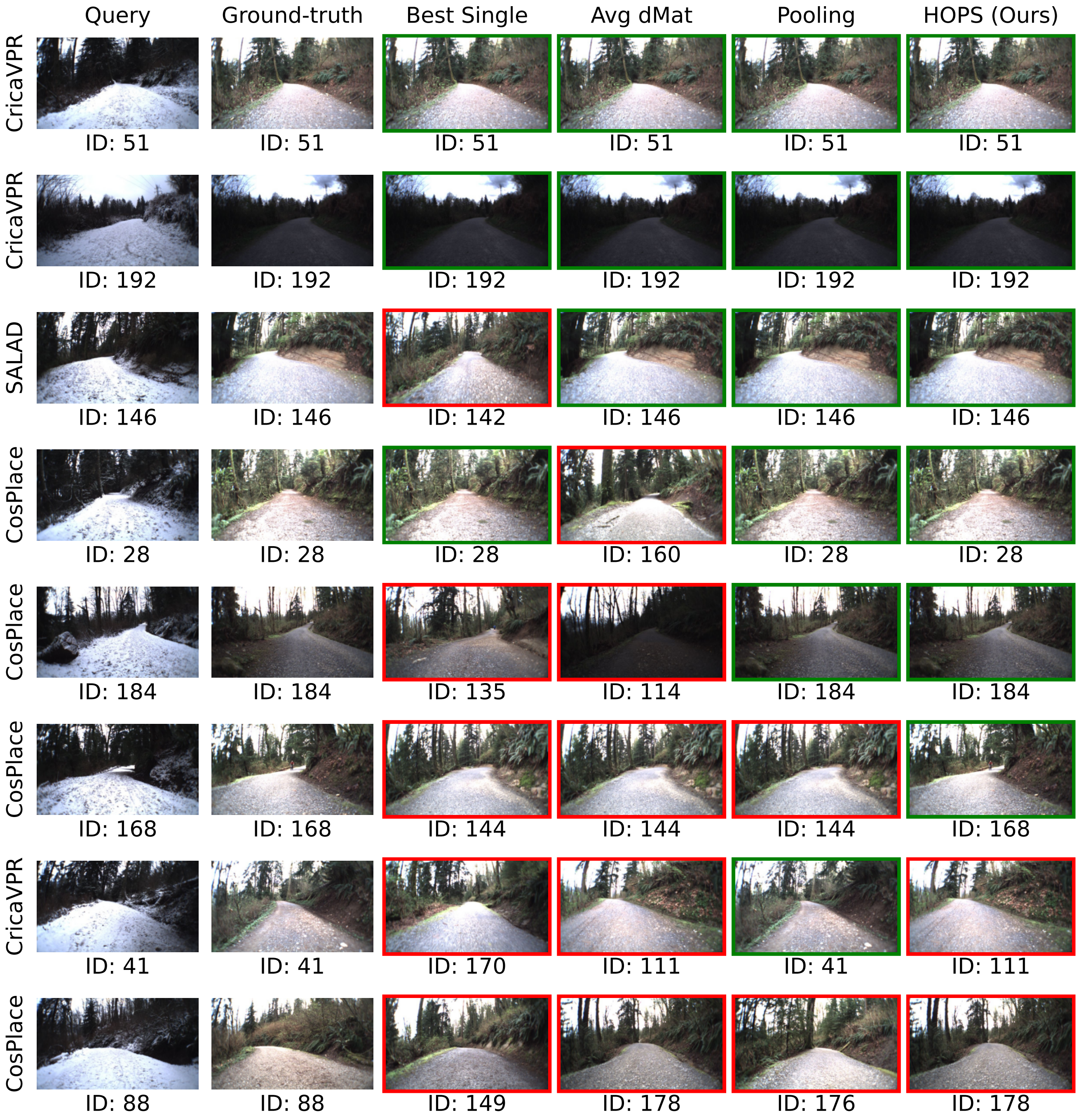

Figure 10. Qualitative results of our method (HOPS) on the SFU-Mountain dataset with January set as the query. For each row, the displayed reference set corresponds to the 'Best Single' reference set for the specific method, shown to the right of the query image. We show three different scenarios: i. Cases where fusing reference sets via HOPS (ours) produces correct matches, similar to other techniques (rows 1 and 2). ii. Cases where HOPS produces correct matches while at least one other method fails (rows 3, 4, 5, and 6). iii. Cases where HOPS retrieves false matches and other methods either succeed or fail (rows 7 and 8). Note that we use a tight ground truth tolerance of ± 1 image for the SFU-Mountain dataset.

References

[1] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Mixvpr: Feature mixing for visual place recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2998–3007, 2023. 8

[2] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Boq: A place is worth a bag of learnable queries. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17794–17803, 2024. 1, 2

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016. 8

[4] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4878–4888, 2022. 2, 3

[5] Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 11080–11090, 2023. 2, 3

[6] Jake Bruce, Jens Wawerla, and Richard Vaughan. The SFU mountain dataset: Semi-structured woodland trails under changing environmental conditions. In *Workshop on Visual Place Recognition in Changing Environments, IEEE International Conference on Robotics and Automation*, 2015. 2

[7] Tobias Fischer and Michael Milford. Event-based visual place recognition with ensembles of temporal windows. *IEEE Robotics and Automation Letters*, 5(4):6924–6931, 2020. 1, 2, 8

[8] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17658–17668, 2024. 6

[9] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 2023. 1

[10] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 2, 6, 8

[11] Peer Neubert and Stefan Schubert. Hyperdimensional computing as a framework for systematic aggregation of image descriptors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16938–16947, 2021. 8

[12] N Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. *Workshop on Long-term Autonomy, IEEE International Conference on Robotics and Automation*, 2013. 2

[13] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 883–890, 2013. 6, 8