2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Visual Place Recognition in Unstructured Driving Environments

Utkarsh Rai[1], Shankar Gangisetty[1], A. H. Abdul Hafez[1], Anbumani Subramanian[1] and C. V. Jawahar[1]

Fig. 1: Illustration of visual place recognition encountering various challenges across the three routes within our unstructured driving VPR dataset. The challenges include occlusions, traffic density changes, viewpoint changes, and variations in illumination. A comparative analysis of these challenges with other datasets is presented in Table I.

***Abstract*— The problem of determining geolocation through visual inputs, known as Visual Place Recognition (VPR), has attracted significant attention in recent years owing to its potential applications in autonomous self-driving systems. The rising interest in these applications poses unique challenges, particularly the necessity for datasets encompassing unstructured environmental conditions to facilitate the development of robust VPR methods. In this paper, we address the VPR challenges by proposing an Indian driving VPR dataset that caters to the semantic diversity of unstructured driving environments like occlusions due to dynamic environments, variations in traffic density, viewpoint variability, and variability in lighting conditions. In unstructured driving environments, GPS signals are unreliable often affecting the vehicle to accurately determine location. To address this challenge, we develop an interactive image-to-image tagging annotation tool to annotate large datasets with ground truth annotations for VPR training. Evaluation of the state-of-the-art methods on our dataset shows a significant performance drop of up to 15%, defeating a large number of standard VPR datasets. We also provide an exhaustive quantitative and qualitative experimental analysis of frontal-view, multi-view, and sequence-matching methods. We believe that our dataset will open new challenges for the VPR research community to build robust models. Project Page: https://cvit.iiit.ac.in/research/projects/cvit-projects/iddvpr**

[1]IIIT, Hyderabad, India utkarsh.rai60@gmail.com, {shankar.gangisetty@ihub-data., ah.abdulhafez@ihub-data., anbumani@, jawahar@}iiit.ac.in

I. Introduction

Visual place recognition is a technique that involves leveraging visual information from the surroundings to identify and distinguish between different places along a route while navigating through various driving scenarios [1]. With the development of autonomous driving technology, self-driving vehicles have moved into real-world complex unstructured driving environments. The unstructured driving environments generally include pickets of shops up to the footpath, irregularly laid sign boards, presence of vegetation at random places, lack of proper lanes, and a very loosely demarcated boundary between a building, footpath, and road. Autonomous driving in unstructured settings requires a suit of abilities that includes perception, planning, and decision-making. One of the most important abilities in this suit is place recognition. Consider a scenario where a vehicle is driving through a city with complex and unstructured road environments (see challenging scenes in Fig. 1). Driving VPR becomes crucial in such situations by enhancing navigation, safety, predictive maneuvering, re-routing, and decision-making for autonomous vehicles and advanced driver assistance systems.

The challenges in VPR have generally been tackled by the

introduction of datasets [2] in structured settings. To drive VPR in unstructured settings there is a need to introduce datasets that tackle the wide gamut of problems associated with such settings. Some of the unstructured driving datasets in literature are IDD [3], IDD-3D [4], Meteor [5], while other datasets include partially unstructured elements such as nuScenes [6], KITTI [7], Waymo [8] and Cityscapes [9] for perception and planning task. Our survey reveals a notable lack of VPR datasets designed specifically for unstructured driving environments [2]. It is important to note that Mapillary Street-level Sequences (MSLS) [10] dataset partially addresses this gap by featuring some scenes within unstructured settings.

Now, let us understand the critical attributes that play pivotal roles within unstructured driving environments, see scenes in Fig. 1 and comparison in Table I: *(i) Long video sequences* - MSLS [10] dataset provide short video sequences of 300 frames which may not capture the complexities of long and diverse driving scenarios. Unstructured environments often involve extended routes with various challenges [3], [4], [5], and short sequences might not adequately represent these situations. *(ii) Presence of dynamic elements* - Unstructured driving environments often involve dynamic elements such as buses, trucks, cars, motorcycles, pedestrians, auto rickshaws, or bicycles. Existing datasets like MSLS [10] and Oxford RobotCar [15] do not adequately simulate or capture the diversity of dynamic objects encountered on the road (see Fig. 2), limiting the model's ability to recognize places in the presence of such elements due to *occlusions* of scenes or locations. *(iii) Adequate environmental variation* - The variability in lighting conditions and weather changes can significantly impact VPR. Datasets like Nordland [13] and Pittsburgh [11] lack comprehensive variations in these environmental factors and may not sufficiently challenge methods to perform robustly in unstructured driving scenarios. *Illumination* scenarios in Fig. 1 depict the discussed challenges. *(iv) Traffic density and unstructured features* - Traffic density can vary widely, ranging from sparse to dense conditions on urban, suburban, and rural roads with unique features like irregular road markings, complex road layouts with intersections and roundabouts, or challenging terrain. Existing datasets like Oxford RobotCar [15] and MSLS [10] do not contain these unstructured features and hence may not effectively address the complexities of place recognition in these diverse driving conditions. *(v) Single-view limitations* - NordLand [13] and MSLS [10] dataset primarily focus on single-view images may not fully capture the multi-perspective challenges associated with unstructured driving. Multi-view datasets like Oxford RobotCar [15] and GSV-Cities [16] are essential for improving recognition performance in complex scenarios.

To address these shortcomings, we introduce the unstructured driving VPR dataset, dubbed IDD-VPR, comprises of, *(i) long sequences* around 10K images for each of the routes (see Table I); *(ii) occlusions* with a large set of dynamic elements (see Fig. 2); *(iii) illumination changes* such as shifting light conditions from day to night and

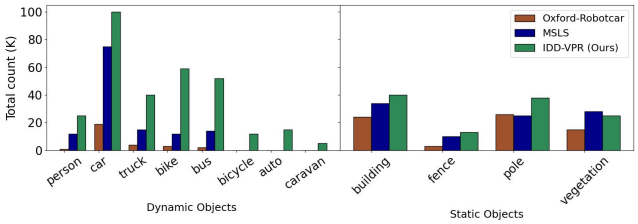

Fig. 2: Comparison of dynamic and static objects in VPR datasets. For evaluation, we take an equal number of images i.e., 20K from MSLS-val, Oxford RobotCar seasons, and our dataset. We see that our dataset has a significantly higher number of dynamic and static agents as well more diverse number of agents.

variations in artificial lighting during nights i.e., vehicle to street lights; *(iv) traffic density changes* representing both low and high traffic density situations in unstructured roads like no-lane marking roads, intersections, and unpaved roads. *(v) viewpoint changes* i.e., multi-view appearance of complex scenarios for improving recognition performance. Fig. 1 depicts samples from the proposed dataset comprising of the above discussed challenges and Table I compares our dataset over the existing VPR dataset addressing the critical attributes of unstructured driving environments. In summary, we make the following contributions:

- We introduce IDD-VPR dataset with the longest coverage of $2,900$ kms in a single city as compared to Oxford RobotCar [15] and St. Lucia [17]. IDD-VPR presents the largest number of images, namely, 34 million images. It is a challenging dataset as it has been captured in diverse unstructured driving environments, comprising long sequences of approximately 10K images per route, and includes *Occlusions*, *Viewpoint Changes*, *Illumination Changes*, and *Traffic Density Changes* (see Fig. 1 and Table I).
- In urban and suburban regions, while capturing data in crowded and unstructured driving environments, GPS signals might be unreliable which often leads to high GPS errors. To address this challenge, we develop an interactive image-to-image tagging annotation tool, integrated with a ranked retrieval algorithm to facilitate fast (query, reference) pair annotation (see Fig. 5). This ensures the consistency of every place being labeled with accurate GPS readings.
- We evaluate the performance of IDD-VPR dataset by comparing against baseline methods such as NetVLAD [18], CosPlace [19], MixVPR [20], Conv-AP [16], and EigenPlaces [21]. The baseline methods performance drops up to 15% on our dataset, defeating a large number of previous datasets. We provide an exhaustive quantitative experimental analysis on frontal-view (see Table II), multi-view (see Table III), and sequence matching (see Table VI), as well as qualitative analysis for image retrieval (see Fig. 7).

II. Visual Place Recognition

Methods. Early VPR research employed handcrafted features such as SIFT and SURF for matching images to

TABLE I: Comparison of datasets for visual place recognition. Total length is the coverage multiplied by the number of times each route was traversed. Time span is from the first recording of a route to the last recording.

Dataset	Location	Environment	Total Length	Time Span	Video Sequence Length (Images)	Total Images	Reference/ Query Images	Occlusions	Single-Multi Viewpoint	Day-Night Illumination	Low-High Traffic
Pitts250k [11]	Pittsburgh	Urban	-	-	-	0.254 M	100 K / 24 K	✗	✓	✗	✗
Tokyo24/7 [12]	Tokyo	Urban	-	-	-	0.174 M	75 K / 315	✗	✓	✓	✓
Nordland [13]	Norway	Train Journey	80 km	1 year	29.7 K	0.028 M	27 K / 2.7 K	✗	✗	✗	✗
SPED [14]	Worldwide*	Urban	-	6 months	-	2.5 M	607 / 607	✗	✓	✓	✗
Oxford RobotCar [15]	Oxford	Urban, Suburban	1,000 km	1 year	9.5 K	20 M	6.5 K / 1 K	✗	✓	✓	✗
Mapillary SLS [10]	Worldwide*	Urban, Suburban	**11,560 km**	7 years	300	0.56 M	19 K / 750	✓	✗	✓	✗
GSV-Cities [16]	Worldwide*	Urban, Suburban	-	**14 years**	-	1.68 M	19 K / 11 K	✓	✓	✓	✗
IDD-VPR (Ours)	**Hyderabad**	**Urban, Suburban**	2,900 km	6 months	10 K	**33.68 M**	30 K / 12 K	✓	✓	✓	✓

*Worldwide relates to multiple cities or multiple countries or multiple continents, but none of these datasets consider majority of Indian driving data.

databases. Moreover, global descriptors like GIS and HoG further broadened VPR scope and applications [1]. The bag of visual words model was introduced to mitigate computational demands, effectively addressing VPR challenges like those in dynamic environments [22].

In the era of deep learning, convolutional neural networks (CNNs) have significantly advanced VPR [23], [24]. Chen *et al.* [25] leveraged the overfeat network [26] with Seq-SLAM [27]. To tackle VPR challenges like scene and viewpoint changes, the methods proposed range from prior-based methods [28], semantic descriptors [29], [30], to cross-domain approaches [31] and environment-specific techniques [32]. Recently hierarchical techniques [33], [34], CNN-based methods [20], [19], [16], [21], and transformer-based approaches [35] have been developed for improved place recognition. NetVLAD [18] introduced an end-to-end trainable architecture specifically for the place recognition task, while CosPlace [19] treated geolocalization as a classification challenge, facilitating place recognition at large scale. Conv-AP [16] improved global descriptor extraction through effective mining techniques, especially beneficial in visually dense environments. MixVPR [20] introduced a method to extract robust descriptors from salient image regions, addressing challenges like illumination. EigenPlaces [21] devised a novel training protocol to enhance model robustness against viewpoint variations.

Datasets. The progression of VPR methods corresponds with the emergence of datasets aimed at tackling the diverse challenges within the domain. Table I presents the comparison of various standard datasets. Early VPR datasets like SPED [14] used 2.5K surveillance cameras for precise ground truth and captured seasonal changes, but lacked in occluded scenes while collected in a low traffic density. Nordland [13], recorded across four seasons from a train-mounted camera, furnishes accurate ground truth and seasonal variations, yet limited in occluded scenes, viewpoint changes, and with no night capture. Oxford RobotCar [15] accommodates over 100 trips on a 10 km route in Oxford and provides highly accurate ground truth. Despite its richness in seasonal changes and a substantial image count of 20 million, it is limited to low-traffic density places and no occluded places.

Two of the popular VPR datasets generated from Google Street View panoramas are Pitts250k [11] and TokyoTM [12]. These datasets feature viewpoint variations and precise GPS coordinates. But they lack scenes with high traffic density and occluded scenes, while Pitts250k also lacks night capture. MSLS [10] is a large dataset covering 30 global cities with viewpoint variations, with limitation in occluded scenes and illumination changes. Most of the MSLS scenes are forward-facing, featuring the frontal-view of the road.

Existing car-mounted VPR datasets like MSLS, Oxford RobotCar, Eynsham [36], and St. Lucia, often feature inaccurate GPS labels and lack coverage across diverse driving conditions. Our dataset addresses all the above challenges and these issues in unstructured driving environments overcoming occlusions, viewpoint, illumination changes, and traffic density, offering improved GPS annotations.

III. Data Capture and Collection

We present the IDD-VPR dataset to push the boundaries of VPR challenges and advance in the state-of-the-art.

Data Capture Platform. To create our dataset, we used the sensor setup and GPS layout as described in IDD-3D [4]. The experimental vehicle drove on three road routes of Hyderabad, India as shown in Fig. 3 to capture various streams of data using 6 camera images at 10 FPS, 1 LiDAR point clouds at 10 FPS, and 1 GPS reading tagged at 1 FPS. Due to dim light at night, we varied the camera exposure times between 5 - 20 ms in comparison to 3 ms for day capture. For synchronization and timestamp of the captured images from different views, we use the open-source ROS libraries. The sensor setup consisted of 6 FLIR Blackfly BFS-U3-32S4C-C cameras of 3.2 MP and 118 FPS, Edmund optics UC series fixed focal length (1×4mm and 4×25mm), 1 Fujinon Fisheye 1.4/1.8mm, 1 Ouster OS1 LiDAR with 64 channel (V) and 1024 channel (H), and 1 G-Star IV BU-353-S4 GPS sensor.

Sensor Calibration. We offer two sets of rectification options tailored to camera lenses with focal lengths of 4mm and 25mm. Additionally, we provide undistortion parameters specifically designed for fish-eye lenses. Since each LiDAR frame corresponds to a corresponding camera frame, we get camera-LiDAR extrinsic parameters using the toolbox proposed in [37].

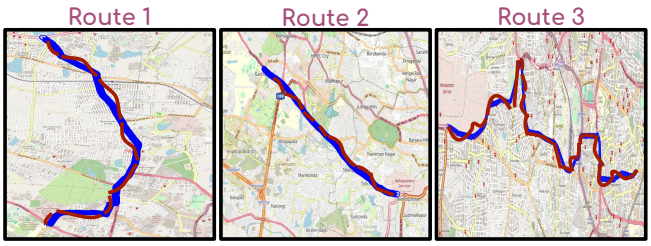

Fig. 3: Data collection map for the three routes. The map displays the actual routes (in blue color) taken and superimposed with maximum GPS drift due to signal loss (dashed lines in red color). This GPS inconsistency required manual correction.

Data Collection. The dataset was collected over 2,900 kms, spanning a period of 6 months from September 2023 to February 2024 across three different routes as shown in Fig. 3, covering 35 trips. The capture vehicle maintained an average speed of 10 meters per second. Data collection along the routes encompassed a wide variety of unstructured driving environments, including lighting conditions (day and night), varying traffic densities (low and high), partial, and no occlusion scenes. We elaborate on each route featuring critical attributes in our dataset within unstructured driving environments as follows:

Route 1 encompasses scenarios capturing structural changes and perceptual aliasing, spanning diverse domains from workplaces and structured settings to single-lane roads with alleyways and unstructured backgrounds. Structural changes are evident in a building undergoing modifications, depicted in the Fig. 1 images labeled as *Occlusions*. A mix of structured and unstructured settings is observed in marketplaces, as depicted in Fig. 1 pairs along Route 1 labeled *Traffic density* and *Viewpoint changes*, respectively. Additionally, this route features narrow alleyways, providing a distinct visual cue compared to other scenarios, as shown in Fig. 1 pair of images labeled *Light*.

Route 2 facilitates the capture of scenarios featuring significant traffic variations and dynamic agents, spanning from peak to off-peak hours. In Fig. 1 pair of images labeled *Traffic density* depicts similar locations with overhead flyovers and marketplaces on the sides. Diverse traffic conditions also result in instances of high occlusion in the recorded scenes. This route additionally includes multiple intersections, flyover passages, and residential apartments, capturing residential settings and crowded scenarios with places of worship nearby, as depicted in Fig. 1 pair of images labeled *Illumination changes*.

Route 3 consists of scenes with high semantic diversity, encompassing landmarks, government buildings, metro stations, large retail stores, and statues. It also exhibits significant variation in *traffic density*, as depicted in Fig. 1. Similar to route 1, this route presents diverse scenes ranging from high-rises to unstructured pickets along single alleyways. Each of these locations features dynamic settings, presented in *Traffic density* and *Illumination changes* pairs, that undergo frequent changes as shown in Fig. 1.

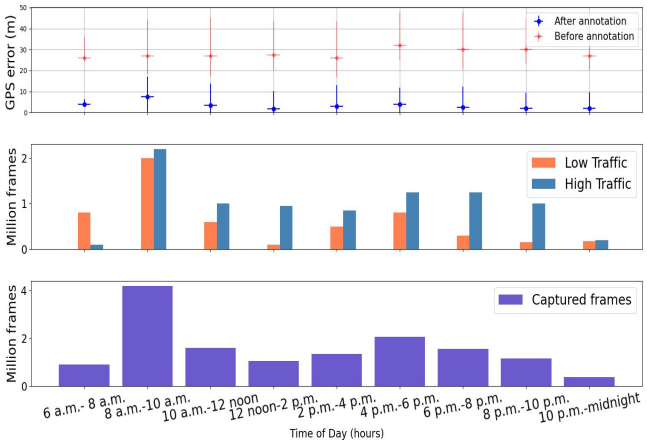

Fig. 4: Data capture distribution per trip and GPS drop. *Bottom plot:* shows the distribution of capture frames in a day, *Middle plot:* highlights the variation of the traffic density during peak and non-peak hours, and *Top plot:* illustrates the GPS errors at different capture times and associated GPS error fixation post annotation.

IV. Data Annotation and Statistics

Annotation. During data capture ensuring consistency and error-free GPS reading for all three route traversals was challenging as shown in Fig. 3. Through our image-to-image tagging annotation process, we ensured the consistency of each location being tagged with the appropriate GPS readings, maintaining a mean error of less than 10 meters. In Table I, we present the release of 33.68 million images. Among these, approximately 6.6 million images have been annotated specifically for scenarios that had significant GPS drifts and fluctuations. In Fig. 4, the top plot illustrates the decrease in mean GPS error compared to the ground truth GPS for the captured samples post-annotation. It's worth noting that the largest GPS drift occurred during night captures, with the highest mean and maximum drift errors being 32 meters and 50 meters, respectively, attributed to crowded and high-traffic scenarios. With the help of our annotation, we were able to accurately label these unstructured driving scenarios.

Annotation Process: We developed an image-to-image matching annotation tool as presented in Fig. 5. For annotation, we involved professional annotators with bespoke training. We ask the annotators to select the front view from the multiple views of the collected samples for annotation. Given a set of 10 *query images* and a *reference image*, the annotators manually choose the most similar *query image* that corresponds to *reference image*. The identified *query image* GPS coordinates are then tagged with the associated *reference image*. To facilitate a faster and better way of annotation, each (query, reference) pair in the annotation tool is ranked according to a scoring mechanism derived from a soft matching technique (employing SIFT descriptor-based key-point matching). The annotators pick the best matching (query, reference) pair based on the ranked sequence from the query panel presented in Fig. 5.

Dataset Features and Statistics. We captured 33.68 million images covering unstructured driving challenges shown for

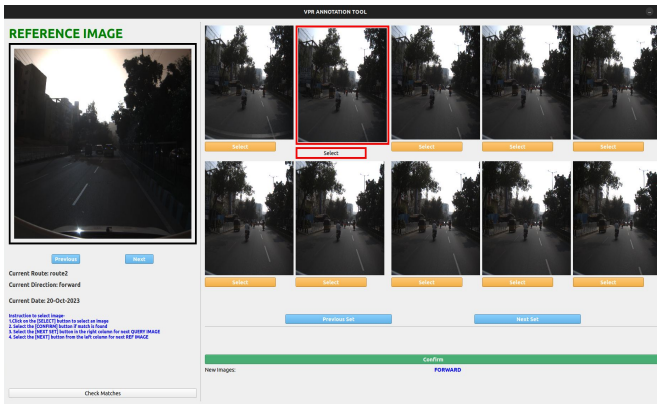

Fig. 5: Image-to-image annotation tool for (query, reference) pair matching by the annotators with GPS tagging.

our dataset in Fig. 1 and important attributes within unstructured driving environments as illustrated in Table I for VPR task. Here, we discuss these challenges and their statistics in our dataset.

For *traffic changes*, we collected 18 million images during high-traffic periods and 15 million images during low-traffic periods, to encompass a wide range of variations in traffic density across peak and off-peak hours as shown in Fig. 4. Our data capture spans all times of the day, encompassing various traffic conditions. We have about 9 million images of high traffic in the morning and 6 million images in the night. In *viewpoint changes*, since all of the images at a particular place are captured by the 6 cameras synchronously, each of the views has 5.5 million images. *Illumination changes* is addressed by distributing our capture throughout the day. We have about 12 million images of evening and night conditions of which nearly 5.5 million images are of the evening (usually having dim lighting conditions), and about 6 million images of early morning covering varying lighting conditions, see Fig. 4 for statistics. In addition, Fig. 6(a) shows capture span along different routes, and Fig. 6(b) presents different samples across changing weathers like overcast, winter, and spring.

V. Experiments

In this section, we discuss the experimental settings, evaluation of proposed IDD-VPR dataset on frontal-view, multi-view, further analysis, and sequence matching for VPR.

A. Experimental Settings

Baseline Methods. We evaluate five baselines on our dataset, namely, NetVLAD [18], CosPlace [19], Conv-AP [16], MixVPR [20], and EigenPlaces [21].These baseline methods are assessed against the unstructured driving environmental challenges introduced in our dataset, namely, *occlusions*, *traffic density*, *viewpoint changes*, and *illumination changes*.
Implementation Details. We implement baseline methods on Intel Xeon E5-2640 v4 processor with four Nvidia RTX 2080 Ti GPUs. The baselines leverage global pooling techniques for single-image retrieval. We evaluate the performance comparison of baselines on our dataset and existing

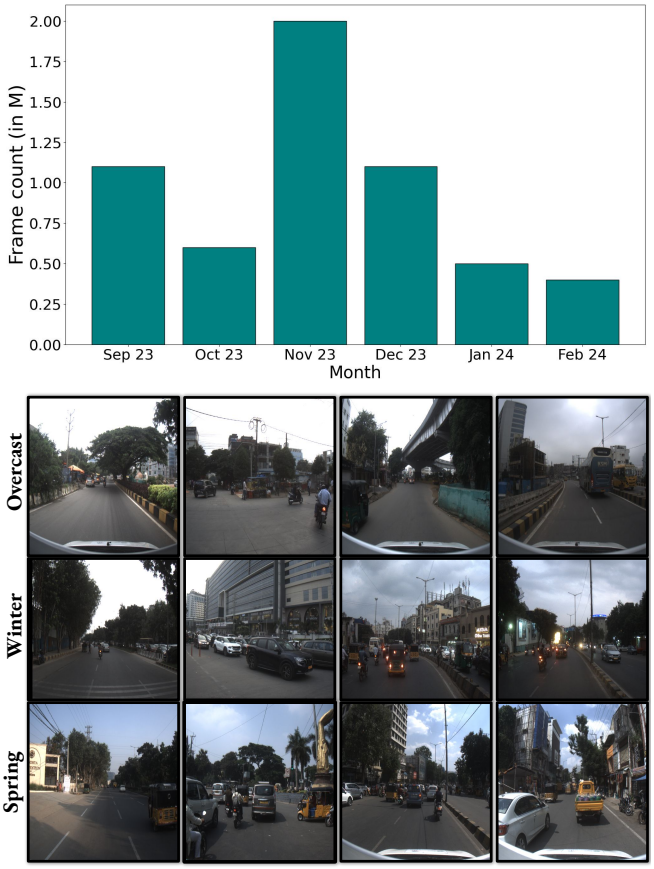

Fig. 6: Data capture span. *Top:* based on months and *Bottom:* diversity of samples encompasses different weather conditions, including overcast (Sep'23, Oct'23), winter (Dec'23, Jan'24), and spring (Feb'24).

datasets with varied descriptor dimensions, and VGG-16 and ResNet-50 backbones. In the frontal- and multi-view experiments, we utilize the same set of 30K *reference images* and *query images* of 12K. The *query images* incorporate variations in traffic density, illumination changes, and front and rear-view scenes. For multi-view experiments, the 12K *query images* are subsampled to include three distinct views, namely, side views (left, right) and rear view of each location.
Evaluation Metric. Recall@k is used as the evaluation metric, that is characterized by the ratio of correctly retrieved queries to the total number of queries. A retrieval is deemed correct if at least one of the top K retrieved images falls within a radius of 25m from the actual position of the query.

B. Frontal-View Place Recognition

In Table II, we report recall@1 experiments on frontal-view datasets, i.e., where the images are frontal road scenes. We observe that recent methods like CosPlace, Conv-AP, MixVPR and EigenPlaces, despite producing compact descriptors are less robust to unstructured driving environmental changes specifically in the context of our dataset. Surprisingly, the earlier method NetVLAD, known for its larger descriptors, significantly underperforms compared to recent methods. Interestingly, no single method stands out as the best performer on our dataset, even though recent methods

TABLE II: Evaluation of baselines on **Frontal-View datasets inclusive of IDD-VPR**: Report overall recall@1, split by utilized backbone and descriptor dimension of 4096-D. See supplementary results video for recall@5/10 with varying descriptor dimensions.

Method	Backbone	Datasets			
		IDD-VPR	Oxford RobotCar	MSLS Val	St. Lucia
NetVLAD [18]	VGG-16	51.9	62.9	59.4	74.8
CosPlace [19]	VGG-16	74.1	88.3	86.9	98.9
Conv-AP [16]	VGG-16	73.2	85.3	86.3	99.3
MixVPR [20]	VGG-16	73.8	87.8	85.2	99.2
EigenPlaces [21]	VGG-16	74.4	87.3	92.7	99.5
NetVLAD [18]	ResNet-50	68.5	81.4	67.8	82.6
CosPlace [19]	ResNet-50	75.6	91.7	88.1	99.6
Conv-AP [16]	ResNet-50	75.2	87.2	87.9	99.3
MixVPR [20]	ResNet-50	**80.8**	92.8	89.5	99.7
EigenPlaces [21]	ResNet-50	76.1	92.8	89.9	99.6

have different characteristics and strengths. MixVPR demonstrates superior performance on frontal-view datasets, with a performance drop of 12%, 8.7%, and 18.9% between our dataset and Oxford RobotCar, MSLS and St. Lucia respectively using ResNet-50 backbone with high-dimensionality descriptors (i.e., 4096-D). Experiments on recall@5 also gives MixVPR to perform superior, with a performance dip of 9.8%, 7.1%, 10.8% (similarly for recall@10 drop of 6.6%, 5.1%, 7.2%) between our dataset and Oxford RobotCar, MSLS and St. Lucia with the same settings. Further details of these results are reported in supplementary video.

In our evaluation of various methods on the IDD-VPR dataset, we observe a notable improvement in recall values with the adoption of denser backbones, specifically ResNet, and larger descriptor dimensions. This trend underscores the demand for more sophisticated representation techniques in handling the intricate and unstructured characteristics of our dataset. The complexity often leads to semantically diverse features appearing closely related in the feature space.

Qualitative Comparison. In Fig. 7, we qualitatively evaluate baseline methods on our dataset. The methods that stand out are those that incorporate advanced training, such as online sample mining and the use of sophisticated loss functions. These approaches enhance VPR training resilience to changes in orientation (as seen in CosPlace and EigenPlaces) and foster more robust feature representations in the face of condition variations (as demonstrated by MixVPR).

C. Multi-View Place Recognition

In Table III, we show recall@1 results on multi-view datasets, i.e., where the reference database is frontal-view while the query orientation can vary across 360°. We demonstrate that existing methods like NetVLAD, CosPlace, Conv-AP, MixVPR, and EigenPlaces, provide lower recall scores even when utilizing compact descriptors, indicating their lack of robustness to diverse unstructured challenges within the context of our dataset. Similar to analysis of frontal-view results, we observe that no method stands out in performance on our dataset, despite these models are learning visual cues from multiple views. Despite not achieving state-of-the-art on our dataset, EigenPlaces has the best overall results, with a performance drop of 10.8%, 10.6%, and 11.3% between ours

TABLE III: Evaluation of baselines on **Multi-View datasets inclusive of IDD-VPR**: Report overall recall@1, split by utilized backbone and descriptor dimension of 4096-D. See supplementary results video for recall@5/10 with varying descriptor dimensions.

Method	Backbone	Datasets			
		IDD-VPR	Pitts30k	Pitts250k	Tokyo 24/7
NetVLAD [18]	VGG-16	63.9	86.9	87.4	71.0
CosPlace [19]	VGG-16	67.1	89.5	90.4	82.9
Conv-AP [16]	VGG-16	66.9	89.5	89.6	62.4
MixVPR [20]	VGG-16	73.8	87.8	91.2	78.4
EigenPlaces [21]	VGG-16	74.4	91.2	92.0	88.4
NetVLAD [18]	ResNet-50	68.9	90.4	87.0	92.4
CosPlace [19]	ResNet-50	69.8	91.5	91.1	84.0
Conv-AP [16]	ResNet-50	69.9	91.8	89.8	65.3
MixVPR [20]	ResNet-50	79.8	92.1	91.6	76.2
EigenPlaces [21]	ResNet-50	**81.7**	92.5	92.3	93.0

and Pitts30k, Pitts250k and Tokyo 24/7 respectively. Experiments on recall@5 also gives EigenPlaces to be superior, with a performance dip of 8%, 9.4%, and 7.3% (similarly for recall@10 drop of 4.6%, 6.1%, 4.4%) between ours and Pitts30k, Pitts250k and Tokyo 24/7 with the same settings. Further details are reported in supplementary video.

The performance disparities among methods on our dataset highlight the unique challenges posed by the multi-view context. EigenPlaces, designed to accommodate for the nuanced multi-view characteristics of VPR datasets, illustrates that the multi-view captures in our dataset result in significant deviations in feature space, which are not adequately addressed by models lacking explicit multi-view training.

D. Further Analysis of Our Dataset Challenges

In this subsection, we go deeper into the richness and diversity of our dataset, presenting qualitative and quantitative analysis of the recognition results. We focus on a curated subset of queries to provide a more in-depth understanding.

Occlusions: In Table IV, we illustrate recall@1 experiments on occlusion changes i.e., where the query images are with and without occlusion. Interestingly, we observe that the performance of the baselines on the occluded scenes drops significantly up to 10% over non-occluded scenes making our dataset more challenging. In Fig. 7, it is evident that a substantial portion of visual cues in certain locations is occluded. Except for MixVPR, other methods struggle and fail to retrieve the correct reference image under these occlusion scenarios.

Viewpoint Changes: In Table V, we report recall@1 experiments on viewpoint changes of our dataset i.e., where the experimental setup is the same as Section V-C, except that the queries vary across sides and rear-view. Interestingly, we observe that rear-view performance dips significantly over side views on all baselines making it more challenging. We also illustrate qualitative analysis in Fig. 7. The challenge associated with frontal/rear traversals arises from the limited visual overlap between these perspectives in our dataset.

Illumination Changes: In Table IV, we demonstrate recall@1 experiments on illumination changes of our dataset i.e., where the query images are day and night. We observe that night light performance drops up to 10% over daylight scenes on all baselines making our dataset more challeng-

Fig. 7: Qualitative comparison of baselines on our dataset. The first column comprises query images of unstructured driving environmental challenges, while the subsequent columns showcase the retrieved images for each of the methods. Green: true positive; Red: false positive.

TABLE IV: Evaluation of baselines on **Traffic Density**, **Illumination Changes**, and **Occlusions** in our dataset: Reported recall@1, split by utilized backbone and descriptor dimension of 4096-D.

Method	Backbone	Traffic Density		Illumination		Occlusion	
		Low	High	Day	Night	w/o	w
NetVLAD [18]	VGG-16	61.9	58.9	66.4	55.3	62.4	55.2
CosPlace [19]	VGG-16	69.4	61.5	72.1	58.2	70.5	61.9
Conv-AP [16]	VGG-16	67.5	62.5	70.1	59.3	68.9	63.1
MixVPR [20]	VGG-16	75.7	69.8	78.3	65.4	76.1	69.4
EigenPlaces [21]	VGG-16	76.4	71.4	78.6	68.2	77.1	70.5
NetVLAD [18]	ResNet-50	69.2	65.4	71.5	60.3	70.1	65.1
CosPlace [19]	ResNet-50	73.2	69.7	76.1	65.8	74.1	65.9
Conv-AP [16]	ResNet-50	74.5	68.6	76.4	63.1	75.1	68.2
MixVPR [20]	ResNet-50	**84.1**	**76.2**	**85.6**	**78.9**	**84.6**	**75.9**
EigenPlaces [21]	ResNet-50	83.4	75.4	83.9	74.2	83.9	75.1

ing. This is attributed to certain scenes being obscured by lighting effects, thereby complicating the process of object identification that adds informative features to the scenes.

Traffic Density: In Table IV, we demonstrate recall@1 experiments on the density of traffic i.e., where the query images are either of low or high traffic. In frontal-view scenarios, we observe that high-traffic conditions significantly impair performance, with a drop of up to 8% compared to low-traffic situations on all baselines. In Fig. 7, we observe that all the baseline methods fail to accurately retrieve the correct reference image. This can be attributed to the fact that high traffic density and the abundance of dynamic yet non-informative agents lead to a decline in performance.

E. Sequence Matching for Place Recognition

Inspired by MSLS [10], we explore sequence matching techniques like seq2seq, seq2im, and im2seq for place recognition in unstructured driving environments. For our experimental analysis, we define the length of each sequence as five frames. In Table VI, we report recall@5 results on sequence matching methods and illustrate the qualitative analysis for our dataset. We take the common backbone and descriptor-dimension i.e., ResNet-50 and 2048-D respectively.

seq2seq: The NetVLAD with MAX sequential pooling method beats the GeM+MAX pooling strategy for *traffic*, *illumination*, and *viewpoint changes* by up to 17%. The outperformance can be attributed to two factors. Firstly, the NetVLAD consistently outperformed GeM across all polling techniques. Secondly, among the NetVLAD methods, employing a max pooling strategy for the embeddings of each image in the sequence, as opposed to average pooling, resulted in more representative descriptors that were more robust to the challenges present in our dataset. In Fig. 8, we observe that NETVLAD with MAX pooling method accurately retrieving the correct reference images across all the unstructured driving environments.

seq2im: The NetVLAD with MODE pooling outperforms GeM+MODE by up to 8% for *traffic density* and *illumination changes*. Similarly, for *viewpoint changes*, NetVLAD with MIN surpasses GeM+MIN by 11.2%. This is because a majority voting for pooling provides us more confidence that a particular image in the database is closer to the query image, as it is the most frequently occurring one.

im2seq: The NetVLAD with MIN performance drops significantly by 20%, 11%, and 28% compared to seq2seq technique for *traffic density*, *illumination*, and *viewpoint changes*. In im2seq, NTVLAD with MIN pooling outperforms GeM+MIN by 4.2%, 18.2%, and 14.1% for all the unstructured driving conditions. This can be attributed to the

TABLE V: Evaluation of baselines on **Viewpoint Changes** in our dataset: Reported recall@1, split by utilized backbone and descriptor dimension of 4096-D.

Method	Backbone	Viewpoint Changes		
		Leftside-View	Rightside-View	Rear-View
NetVLAD [18]	VGG-16	62.4	61.4	57.2
CosPlace [19]	VGG-16	67.1	66.4	62.3
Conv-AP [16]	VGG-16	67.9	66.4	62.9
MixVPR [20]	VGG-16	72.1	71.3	69.3
EigenPlaces [21]	VGG-16	74.2	75.3	73.6
NetVLAD [18]	ResNet-50	65.4	67.4	62.2
CosPlace [19]	ResNet-50	69.1	69.4	66.3
Conv-AP [16]	ResNet-50	68.9	69.3	66.1
MixVPR [20]	ResNet-50	80.1	80.3	73.3
EigenPlaces [21]	ResNet-50	**82.2**	**83.4**	**76.9**

Fig. 8: Qualitative comparison of baselines for *seq2seq* matching on our dataset. The first column comprises query images of unstructured driving environmental challenges, while the subsequent columns showcase the retrieved images for each of the seq2seq methods. Green: true positive; Red: false positive.

TABLE VI: Evaluation of ***seq2seq/seq2im/im2seq methods*** on **Traffic Density**, **Illumination Changes**, and **Viewpoint Changes** in our dataset: Reported recall@5, using ResNet-50 backbone, training dataset of Pitts250k, and descriptor dimension of 2048-D. The provided queries correspond to conditions of low traffic, daytime, and a frontal viewpoint, respectively.

Matching	Method	Low/High Traffic	Day/Night Illumination	Front/Rear Viewpoint
seq2seq	NetVLAD+MAX	**71.2**	**68.5**	**72.3**
	NetVLAD+AVG	60.1	59.1	63.4
	GeM+MAX	53.4	51.8	55.6
	GeM+AVG	55.2	57.9	52.5
seq2im	NetVLAD+MIN	61.1	59.3	**64.6**
	NetVLAD+MODE	**65.7**	**60.8**	58.5
	GeM+MIN	58.1	57.2	53.4
	GeM+MODE	57.7	54.4	51.2
im2seq	NetVLAD+MIN	**41.1**	**47.9**	**36.6**
	GeM+MIN	36.9	29.7	22.5

image-to-sequence approach, where we match query images to all frames in the reference sequences and then select the sequence with the closest frames.

VI. Conclusions

In this work, we presented a large dataset IDD-VPR, tailored for unstructured driving environmental challenges including occlusions, variations in traffic density, viewpoint changes, and variability in illumination conditions. We develop an interactive image-to-image annotation tool for accurate GPS tagging to handle GPS-denied environments. Extensive experiments with state-of-the-art baselines on our dataset demonstrate its complexity through quantitative and qualitative assessments. These assessments reveal performance degradation compared to existing datasets across frontal view, multi-view, and sequence matching tests. Through this dataset and future releases, we shall extend to broaden the scope of applications and enhance the integration of VPR into self-driving autonomous systems. This will improve their ability to navigate, localize, and interact safely and efficiently with diverse unstructured environments in various geographic locations.

References

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *Transactions on Robotics*, 2015.

[2] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognition*, 2021.

[3] G. Varma, A. Subramanian, A. M. Namboodiri, M. Chandraker, and C. V. Jawahar, "IDD: a dataset for exploring problems of autonomous navigation in unconstrained environments," in *WACV*, 2018.

[4] S. Dokania, A. H. A. Hafez, A. Subramanian, M. Chandraker, and C. V. Jawahar, "IDD-3D: Indian driving dataset for 3d unstructured road scenes," in *WACV*, 2023.

[5] R. Chandra, X. Wang, M. Mahajan, R. Kala, R. Palugulla, C. Naidu, A. Jain, and D. Manocha, "METEOR: A dense, heterogeneous, and unstructured traffic dataset with rare behaviors," in *ICRA*, 2023.

[6] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[7] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *IJRR*, 2013.

[8] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, 2020.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.

[10] F. Warburg, S. Hauberg, M. López-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *CVPR*, 2020.

[11] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla, "Visual place recognition with repetitive structures," *TPAMI*, 2015.

[12] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 Place recognition by view synthesis," *TPAMI*, 2018.

[13] D. Olid, J. M. Fácil, and J. Civera, "Single-view place recognition under seasonal changes," *CoRR*, vol. abs/1808.06516, 2018.

[14] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. D. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *ICRA*, 2017.

[15] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *IJRR*, 2017.

[16] A. Ali-bey, B. Chaib-draa, and P. Giguère, "Gsv-cities: Toward appropriate supervised visual place recognition," *Neurocomputing*, 2022.

[17] A. Glover, W. P. Maddern, M. Milford, and G. F. Wyeth, "FAB-MAP + RatSLAM: Appearance-based SLAM for multiple times of day," in *ICRA*, 2010.

[18] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," *TPAMI*, 2018.

[19] G. M. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *CVPR*, 2022.

[20] A. Alibey, B. Chaib-draa, and P. Giguère, "MixVPR: Feature mixing for visual place recognition," in *WACV*, 2023.

[21] G. M. Berton, G. Trivigno, B. Caputo, and C. Masone, "EigenPlaces: Training viewpoint robust models for visual place recognition," in *ICCV*, 2023.

[22] A. H. A. Hafez, M. Arora, K. M. Krishna, and C. V. Jawahar, "Learning multiple experiences useful visual features for active maps localization in crowded environments," *Advanced Robotics*, 2016.

[23] L. G. Camara, T. Pivoňka, M. Jílek, C. Gäbert, K. Košnar, and L. Přeučil, "Accurate and robust teach and repeat navigation by visual place recognition: A cnn approach," in *IROS*, 2020.

[24] H. Wang, C. Wang, and L. Xie, "Online visual place recognition via saliency re-identification," in *IROS*, 2020.

[25] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," *ArXiv*, vol. abs/1411.1509, 2014.

[26] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," in *ICLR*, 2014.

[27] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *ICRA*, 2012.

[28] L. Tang, Y. Wang, Q. Luo, X. Ding, and R. Xiong, "Adversarial feature disentanglement for place recognition across changing appearance," in *ICRA*, 2020.

[29] S. Garg, N. Sünderhauf, and M. Milford, "LoST? Appearance-invariant place recognition for opposite viewpoints using visual semantics," in *Robotics: Science and Systems*, 2018.

[30] A. Gawel, C. D. Don, R. Siegwart, J. I. Nieto, and C. Cadena, "X-View: Graph-based semantic multiview localization," *RA-L*, 2018.

[31] S. Ibrahimi, N. van Noord, T. Alpherts, and M. Worring, "Inside out visual place recognition," in *BMVC*, 2021.

[32] N. V. Keetha, M. Milford, and S. Garg, "A hierarchical dual model of environment- and place-specific utility for visual place recognition," *RA-L*, 2021.

[33] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "TransVPR: Transformer-based place recognition with multi-level attention aggregation," in *CVPR*, 2022.

[34] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, "R2former: Unified retrieval and reranking transformer for place recognition," in *CVPR*, 2023.

[35] N. V. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. A. Scherer, K. M. Krishna, and S. Garg, "AnyLoc: Towards universal visual place recognition," *RA-L*, 2024.

[36] M. J. Cummins and P. Newman, "Highly scalable appearance-only SLAM-FAB-MAP 2.0," in *Robotics: Science and Systems*, 2009.

[37] K. Koide, S. Oishi, M. Yokozuka, and A. Banno, "General, single-shot, target-less, and automatic LiDAR-Camera extrinsic calibration toolbox," in *ICRA*, 2023.