NYU-VPR: Long-Term Visual Place Recognition Benchmark with View Direction and Data Anonymization Influences

Diwei Sheng*, Yuxiang Chai*, Xinru Li, Chen Feng†, Jianzhe Lin, Claudio Silva, John-Ross Rizzo

***Abstract*— Visual place recognition (VPR) is critical in not only localization and mapping for autonomous driving vehicles, but also assistive navigation for the visually impaired population. To enable a long-term VPR system on a large scale, several challenges need to be addressed. First, different applications could require different image view directions, such as front views for self-driving cars while side views for the low vision people. Second, VPR in metropolitan scenes can often cause privacy concerns due to the imaging of pedestrian and vehicle identity information, calling for the need of data anonymization before VPR queries and database construction. Both factors could lead to VPR performance variations that are not well understood yet. To study their influences, we present the NYU-VPR dataset that contains more than 200,000 images over a 2km×2km area near the New York University campus, taken within the whole year of 2016. Our benchmark results on several popular VPR algorithms show that side views are significantly more challenging for current VPR methods while the influence of data anonymization is almost negligible.**

I. Introduction

Visual place recognition (VPR) is the process of retrieving the most similar images for a query one from a database of images with known camera poses, which is often used for loop closing in mapping, localization, and navigation. It relies on representing an image as a global feature vector which describes the portion of the image appearance that is most relevant to its capturing pose. Its applications range from autonomous driving for vehicles, to assistive navigation for the visually impaired people, especially in busy and crowded metropolitan areas where GPS could suffer from the "urban canyon" problem when satellite signals are blocked or multi-reflected to cause large localization errors.

A reliable large-scale and long-term VPR system has to address several challenges. The first one is to choose a proper image view direction. In a self-driving car scenario, using front-view images from a dash-mounted camera whose view direction is parallel to the driving/street direction is almost a default choice. In fact, most VPR methods have been investigated and evaluated under such front-view conditions, where features of roads, shapes of skylines, and textures of the roadside buildings can all contribute to describing and discriminating various image locations.

Yet not all downstream applications prefer front-view images. For example, a person with low vision might need image-based wearable navigation assistance to find the entrance of a particular shop on the street. The aforementioned front-view images typically contain more than half of the pixels on the road and the sky while the remaining pixels on street sides that are far away from the image capturing locations. Contrarily, the side view offers fronto-parallel images of buildings along streets, which stores enough information required for such an application scenario.

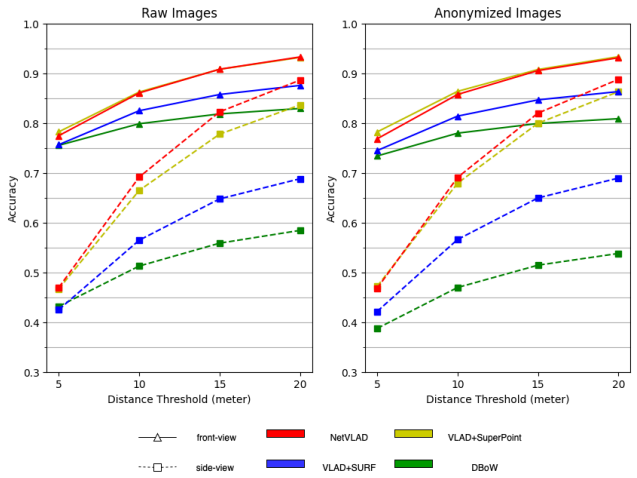


Fig. 1. Top-1 VPR retrieval accuracy of 4 baseline algorithms under different view directions and data anonymization.

However, currently, there is a lack of datasets that could evaluate VPR methods specifically on side-view images in comparison with front-view ones. Many datasets contain no side-view images or mix them with front-view images without explicit labels (see Table I). Moreover, as far as we know, there is no systematic comparison of VPR performance between images from the two view directions. Thus, the following questions remain unclear although one might have correct intuitions: *are side-view images more challenging for VPR than front-view ones? And if so, how much is the performance difference?*

Another challenge for large-scale VPR systems is data privacy, receiving an increasing attention from the community [7]. Building such systems in large metropolitan areas requires collecting images for a long term, inevitably creating concerns of both violating the privacy of the identity information of individual pedestrians and vehicles, and potentially even leaking their spatial-temporal trajectories. Unlike the privacy-preserving technique in [7] that still operates on raw images, another way could be directly anonymizing the images by wiping out all the identity-related pixels (see Figure 3). However, this again brings some unanswered questions for VPR: *would such data anonymization significantly affects the performance of existing VPR algorithms? If so, does it increase or decrease the VPR accuracy/robustness?*

New York University, Brooklyn, NY 11201, USA
* indicates equal contributions.
†Chen Feng is the corresponding author. cfeng@nyu.edu

TABLE I

COMPARISON OF MAJOR PUBLIC OUTDOOR VPR DATASETS WITH NYU-VPR.

Dataset	side-view	side-view-label	dynamic-object	crowded-area	anonymization	seasonal-changes	#images
Nordland [1]	✗	-	✗	✗	✗	✓	28,865
VPRiCE 2015 [2]	✗	-	✓	✗	✗	✗	7,778
Tokyo 24/7 [3]	✓	✗	✓	✓	face-only	✗	76,000
Pittsburgh [4]	✓	✗	✓	✗	✗	✓	254,064
KITTI raw [5]	✗	-	✓	✗	✗	✗	12,919
KAIST [6]	✗	-	✓	✗	✗	✗	105,000
NYU-VPR(ours)	✓	✓	✓	✓	✓	✓	201,790

This paper aims at filling the gaps by introducing a new VPR dataset and benchmark. This is a year-long dataset captured outdoors by vehicles with front-view and side-view cameras traveling in a metropolitan area recording more than 200,000 GPS-tagged images. This allows us to answer the questions we raised before by comparing the results under various conditions.

The contributions of this paper are as follows:

- We present NYU-VPR, a unique large-scale, year-long, outdoor VPR benchmark dataset containing both front-view and side-view GPS-tagged images taken at different lighting conditions with seasonal and appearance changes in a busy and crowded urban area of New York City. This dataset and our benchmark code will be released for educational and research purposes.
- We benchmark the performance of several popular VPR algorithms with a focus on the influence of image view directions. As far as we know, this is the first work to systematically demonstrate the significant challenge of VPR with side-view images.
- We anonymize the identify information in this dataset by removing pixels of both pedestrians and vehicles to address the privacy concerns of large-scale VPR in urban scenes. This is also the first result to show that all the benchmarked VPR algorithms are only marginally affected by this anonymization.

II. RELATED WORK

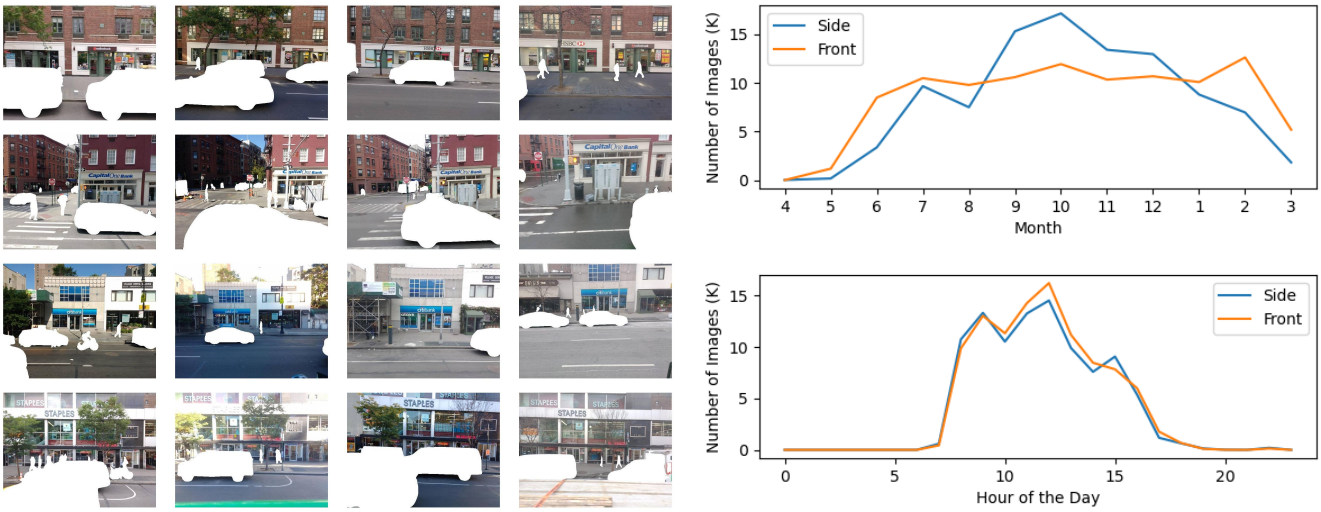
Because NYU-VPR contains only outdoor images used for visual place recognition, we review publicly available datasets that have similar characteristics. The main differences between those datasets and our proposed dataset are summarized in Table I.

Side-view and side-view label: In recent years there has been substantial growth in the number of visual place recognition datasets in the urban areas. However, most datasets only contain front-view images, gathered by cameras on the front and back of cars [1–6]. The side-view images featuring the storefront and sidewalk are not included. Few datasets include the side-view images in addition to the front-view images. For example, the images in Tokyo 24/7 dataset were gathered by pedestrians' phones and featured both front-view and side-view images [3]. The images in Pittsburgh dataset were perspective images generated from Google Street View panoramas of the Pittsburgh area [4]. But those datasets do not label the images as front-view or side-view. Thus no work can use those datasets to compare the visual place recognition results on the side-view images versus those on the front-view images. NYU-VPR contains images labeled as front-view and images labeled as side-view. We focus on evaluating the visual place recognition algorithms in both categories. We compare the results of algorithms on the side-view images versus the results on the front-view images in order to analyze the influence of view direction on the long-term visual place recognition.

Dynamic objects, crowded-area and anonymization: Dynamic objects such as pedestrians and vehicles in images may affect the performance of visual place recognition due to the changing appearance at the same place. Besides, the presence of pedestrians and vehicles in publicly available datasets may raise privacy issues if images are not anonymized. There are few appearances of dynamic objects in datasets of images gathered in suburban areas. For example, Nordland is a dataset of images taken on the train on a railway line between the cities of Trondheim and Bodø [1]. In contrast, dynamic objects appears much more frequently in the datasets featuring urban areas [2–6]. We define crowded areas as metropolitan areas such as New York City and Tokyo that has a high population density and is crowded with pedestrians and vehicles. In Table I, images from Tokyo 24/7 and NYU-VPR are gathered in crowded areas. Anonymization is needed on those datasets for privacy protection. Tokyo 24/7 only applied face redaction on pedestrians [3]. We use MSeg [8] instead of face redaction to replace pedestrians and vehicles with white pixels.

Seasonal changes: Matching images that are taken at the same location in different seasons is crucial for long-term visual place recognition. This is because objects on images change with the seasons: new storefront, trees withering, constructions finished, etc. In Table I, Pittsburgh [4] includes images in different seasons but few image locations are visited in all four seasons. Nordland and NYU-VPR includes images in four seasons for most locations. For Nordland, every location is visited once every season. For NYU-VPR, most location is visited more than once every season as shown by the time distribution of images in Fig. 2(b) and space distribution of images in Fig. 4(f) and Fig. 4(e). So NYU-VPR can be used to evaluate long-term visual place recognition for the influence of seasonal changes.

Baseline methods: Current VPR methods can be roughly grouped into three categories: deep-learning methods, non-deep-learning methods, and methods that only use deep



(a) Images of 4 locations (row) at 4 months (3 months per column).

(b) Month/hour distributions of images.

Fig. 2. Dataset visualization of NYU-VPR with respect to the image capturing time.

learning descriptors. We select methods in each of the three categories. Deep learning methods [9–12] use a convolutional neural networks (CNN) and train CNN in an end-to-end manner directly. We select NetVLAD [9] and PoseNet [10] in this group. For non-deep-learning methods [3, 13–15], two classical ones are bag-of-words (BoW) model and Vector of Locally Aggregated Descriptors (VLAD). In this group, we choose DBoW+ORB [13, 16], which was used in the popular ORB-SLAM for loop closing [17], and VLAD+SURF [14, 18], which was used in [19]. Methods that only use deep learning descriptors take advantage of deep nets' ability to detect a richer set of key points, such as SuperPoint [20] which was used in the first place at ECCV 2020 long-term visual localization under changing conditions workshop [21]. So we also choose VLAD+SuperPoint.

III. The NYU-VPR Dataset

Our dataset is named NYU-VPR. It is composed of images recorded in Manhattan, New York from April 2016 to March 2017. The images were recorded by cameras installed on the front, back, and side parts of taxis with auto-exposure. The dataset contains both side-view images and front-view images. There are 100,500 side-view and 101,290 front-view images, each with a 640×480 resolution. On the basis of raw images, we use MSeg [8], a semantic segmentation method, to replace moving objects such as people and cars with white pixels. Fig. 3 compares anonymized and raw images.

Fig. 3. Raw images vs. Anonymized images.

The images were recorded on streets around Washington Square Park. The trajectories of the locations where the images were recorded are shown in Fig. 4(d). Since the cameras were placed on taxis and taxis' routes were random, the frequencies of locations where the images were taken are different. The frequencies of the locations where the side-view and front-view images were recorded are shown in Fig. 4(f) and Fig. 4(e) respectively.

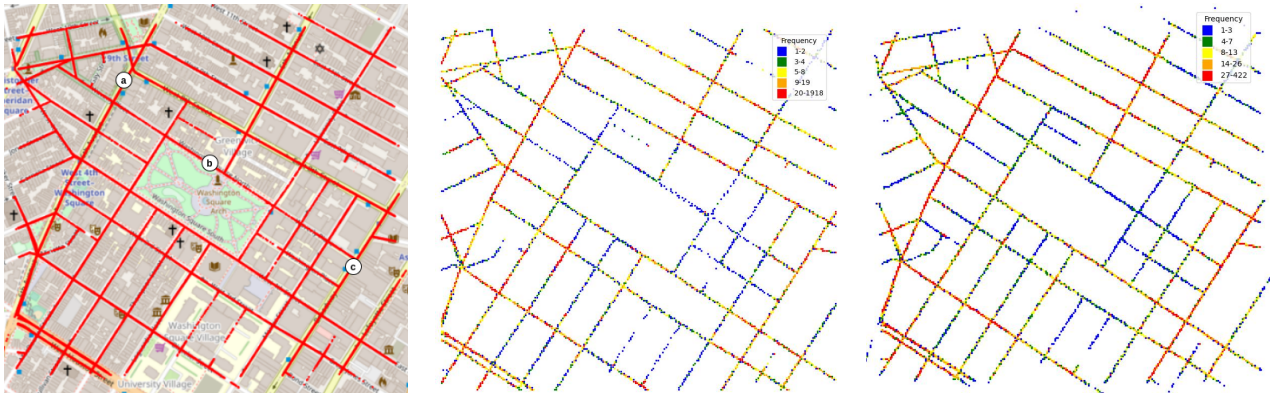
The dataset contains captured images for one year long from April 2016 to March 2017. Fig. 2(b) shows the time distribution. Our dataset includes all four seasons. Therefore, it contains weather changes, illumination changes, vegetation changes, and road construction changes. As shown in Fig. 2(a), we can see image changes at the same location as the season changes.

Difficulty Level: We assign each side-view query image a difficulty level of easy, medium, or hard. First, we extract SIFT [22] features for each image. Then for each query image, we find the top-8 closest side-view training images by GPS coordinates. The query image and its top-8 closest images form eight image pairs. We use RANSAC to compute a fundamental matrix and the number of inliers for each pair of images. We use three intervals to measure the difficulty level of matching each pair based on each pair's number of inliers points: 0-19 (hard), 20-80 (medium), >80 (easy). The interval values are determined by artificially viewing the image pairs and checking the similarity of the image pairs. The difficulty level of each side-view query image is the most common difficulty level of its eight pairs.

Uniqueness: Comparing to front-view images where sky and road surfaces occupy large areas, side-view images provide more useful information about the location such as explicit shop signs and metro entrances. The advantage of anonymization is the privacy protection of people and cars. And in the meantime, anonymized images provide algorithms static and environment-only information, getting rid of moving and interfering objects.

Challenges: There are several challenges in our dataset. Because our dataset is one-year long, the images taken at the same location may have several differences artificially and naturally. First, Fig. 4(a) shows two images. The left one was taken in October 2016 with sideway constructions

(a) During vs. After construction. (b) Summer vs. Winter. (c) Without vs. With motion blur.



(d) GPS locations of images in NYU-VPR. (e) Frequency of front-view images. (f) Frequency of side-view images.

Fig. 4. Dataset visualization of NYU-VPR with respect to the image capturing location. The locations of (a)-(c) are highlighted in (d).

and the right image was taken in December 2016 after the construction. The construction does not occupy much space here, but constructions may cover the whole image at some location. Second, different seasons cause different appearances at the same location. Fig. 4(b) shows two images. The left one was taken in summer, July 2016, and the right one was taken in winter, January 2017. According to Fig. 4(b), the vegetation in Washington Square Park had changed a lot and snow was covering the ground in winter. Furthermore, if the vehicle was moving fast, the images taken by the vehicle will be blurry. Fig. 4(c) shows two images. The left one is not blurry but the right one is blurry. Although two images were taken at the same location, the blurry one will cause more difficulty during the retrieval.

IV. Benchmark Experiments

A. Settings

We selected five classical as well as state-of-the-art descriptors and methods for evaluation of visual place recognition performance on the NYU-VPR dataset.

VLAD+SURF: We use Vector of Locally Aggregated Descriptors(VLAD) [14] to aggregate speeded up robust features (SURF) [18] descriptors for image retrieval. Through experiments, we find the optimal cluster number is 32 within 8, 16, 32, and 64, by using MiniBatchKmeans with batch size at 5000. This cluster number gives high accuracy and acceptable training time. The training of 77608 images took about 8 hours on CPU with 64 GB available memory.

VLAD+SuperPoint: We use SuperPoint model pre-trained on MS-COCO generic image dataset [20]. We use nVidia RTX 2080S to extract SuperPoint features. Then we use VLAD to aggregate SuperPoint descriptors for image retrieval. We set the cluster number at 32, just as we do in VLAD+SURF, by using MiniBatchKmeans with batch size at 100. Notice the dimension of SuperPoint descriptors is larger than the dimension of SURF descriptors. The training of 77608 images took about 20 hours on CPU and GPU with 64 GB available memory.

NetVLAD: We directly use the pre-trained model weight for 30 epochs on Pittsburgh-250k datasets [4] to complete our testing. For the hardware, the CPU we adapt is Intel® Core™ i7-8700k, and the GPU we use is NVIDIA GEFORCE GTX 1080 TI. We first complete an initial clustering on training data to find out the centroids used for the testing process. The input testing data with the extracted deep feature are assigned to different clusters afterward. The batch size during testing is 24.

PoseNet: We use PoseNet model with ResNet34 as the base architecture [10]. For training, PoseNet requires the Cartesian coordinates of images as input besides images themselves. So we gather latitude and longitude information of training images from the camera. We convert latitude and longitude to universal transverse mercator (UTM) coordinates to improve the accuracy of PoseNet's estimation of images' relative position. We use images with normalized UTM coordinates as the input to PoseNet. The GPU we use is NVIDIA GEFORCE RTX 2080S. The batch size during training is 32. For testing, PoseNet outputs the estimation of normalized UTM coordinates of the query images, which are used for evaluation.

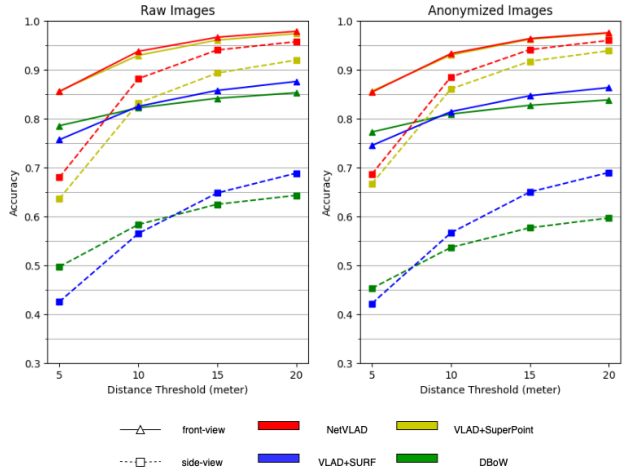
DBoW: We use Distributed Bag of Words (DBoW) [23] model. We choose Oriented FAST and Rotated BRIEF (ORB) [16] descriptors for representing features. We use DBoW to generate a vocabulary constructed by ORB descriptors of training and test images. For testing, We generate the top-5 retrieval images by using DBoW3 to generate a score between each training and test images and selecting the top-5 scores for each test image. We run the testing process with

multi-thread for efficiency.

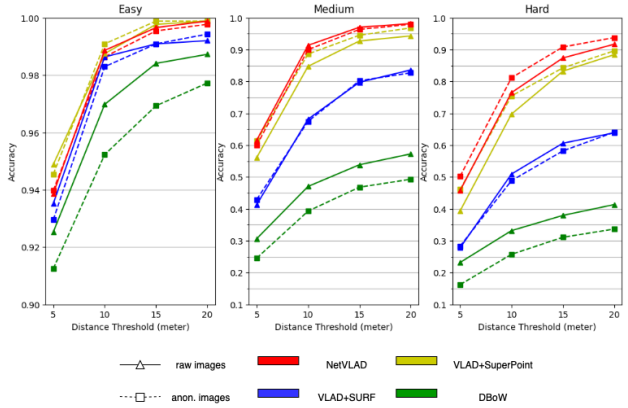
Dataset: We randomly split both front-view and side-view images into training, validation and testing sections by 80%, 5%, and 15% respectively. For each view direction, both anonymized and raw images share the same split result. All images are resized to 640×480. We also use Python module utm to convert GPS coordinates to UTM coordinates for more precise distance calculation between two locations.

Evaluation: Just as in VLASE: Vehicle Localization by Aggregating Semantic Edges [19], we measure both top-1 and top-5 retrieval accuracy under four distance thresholds (5, 10, 15, 20 meters). if any of the top-k retrieval images are within the range of the distance from the query image, we count it as a successful retrieval.

B. Results



(a) Top-5 retrieval results.



(b) Top-5 retrieval results in terms of difficulty level.

Fig. 5. Main results of the benchmark.

Figure 5(a) shows our main results. The result is mainly focusing on the performance of anonymized side-view dataset, non-anonymized side-view dataset, anonymized front-view dataset, and non-anonymized front-view dataset. The comparison among these four experiments shows the performance of the methods on both side-view and front-view datasets as well as the influence of anonymization and view direction.

Performance: Fig. 1 and Fig. 5(a) shows the result of our experiments. Obviously, the result of the top 5 retrieval image accuracy is higher than the top 1 retrieval image accuracy, with an average 10% higher. And when using VLAD to aggregate descriptors, the result of SuperPoint descriptors is much more accurate than the result of SURF descriptors. In both two figures, the accuracy of NetVLAD is the highest, followed by VLAD+SuperPoint and VLAD+SURF. The last is the DBoW method. We may attribute the low accuracy of DBoW to the unsustainability of ORB features. PoseNet outputs a GPS coordinate and using that coordinate we can find the closest top 1 retrieval image, and through experiments, the accuracy of PoseNet is 15.3% and 37.5% when the distance threshold is 5 and 10 meters respectively. Due to the low performance of PoseNet, we omit it in other experiments and not to show the results. We also calculate the accuracy in terms of difficulty level as mentioned before. Fig. 5(b) shows the result of top 5 retrieval in different difficulty level. The visual result is shown in Fig 6. Clearly, we can see VLAD+SuperPoint has better performance than VLAD+SURF and DBoW.

Anonymization: From Fig. 1, we can draw the conclusion that the anonymization does not have a large impact on the visual place recognition result, either of front-view dataset or side-view dataset. The anonymization, however, has little influence on the accuracy of some methods. For example, VLAD+SuperPoint gets 1.1% increase on average, while DBoW and VLAD+SURF have around 2.1% and 3.4% decrease on average respectively. Therefore, when doing experiments on visual place recognition, we can anonymize raw images to protect the privacy of people and cars.

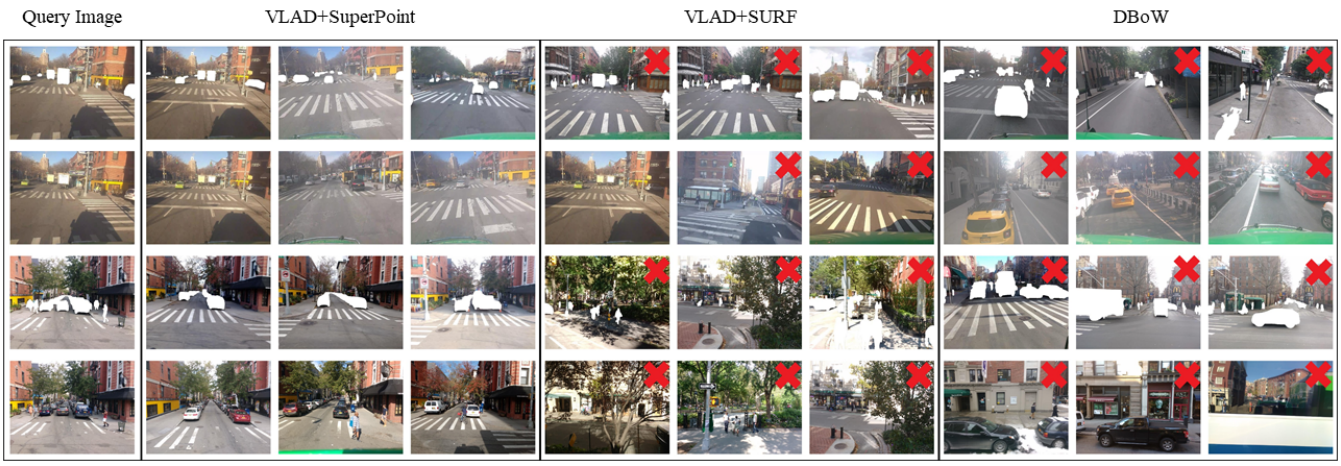
View Direction: View direction does have a conspicuous influence on retrieval accuracy. As we can see in Fig. 1, the accuracies calculated from front-view images are higher than those from side-view images. This phenomenon happens in every method, so we can draw the conclusion that using front-view images is better than using side-view images when doing visual place recognition experiments.

V. Conclusions

After the large-scale experiments and analysis, we can finally answer our questions with confidence. Are side-view images more challenging for VPR than front-view ones? *Yes, and the performance drops of all VPR methods are significant*, although the dataset has no significant spatial/temporal differences on the distribution of images captured from the two view directions. Would our data anonymization significantly affect the performance of existing VPR algorithms? *No, and for some methods, the anonymization could even bring marginal improvements*, potentially due to the removal of those VPR noises. Our future work includes benchmarking more VPR methods and with geometric verification.

Acknowledgments

The research is supported by C2SMART and USDOT Award #69A3551747124.



(a) location1

(b) location2

(c) location3

Fig. 6. We choose VLAD+SuperPoint, VLAD+SURF, DBOW, because the result of NetVLAD is very similar to that of VLAD+SuperPoint. We randomly picked three locations. In each subfigure, the first row is the result of an anonymized query front-view image. The second row is the result of the corresponding raw front-view image. The third row is the result of the query side-view image which was taken at the same location as the query front-view image. The last row is the result of the corresponding raw side-view image. The red cross means the location of the retrieval image is not in the distance threshold (10 meters). We show the top 3 retrievals for each method here.

References

[1] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," in *Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA)*, 2013, p. 2013. 2

[2] "The VPRiCE Challenge 2015 – Visual Place Recognition in Changing Environments - Public - Confluence." [Online]. Available: https://roboticvision.atlassian.net/wiki/spaces/PUB/pages/14188617/The+VPRiCE+Challenge+2015+Visual+Place+Recognition+in+Changing+Environments 2

[3] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1808–1817. 2, 3

[4] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *CVPR*, 2013. 2, 4

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013. 2

[6] Y. Choi, N. Kim, K. Park, S. Hwang, J. S. Yoon, Y. In, and I. Kweon, "All-day visual place recognition: Benchmark dataset and baseline," in *Workshop on Visual Place Recognition in Changing Environments*, 06 2015. 2

[7] P. Speciale, J. L. Schonberger, S. B. Kang, S. N. Sinha, and M. Pollefeys, "Privacy preserving image-based localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5493–5503. 1

[8] J. Lambert, Z. Liu, O. Sener, J. Hays, and V. Koltun, "MSeg: A composite dataset for multi-domain semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[9] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307. 3

[10] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946. 3, 4

[11] M. Chancán, L. Hernandez-Nunez, A. Narendra, A. B. Barron, and M. Milford, "A hybrid compact neural architecture for visual place recognition," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 993–1000, April 2020.

[12] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3223–3230. 3

[13] D. Gálvez-López and J. D. Tardós, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, October 2012. 3

[14] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3304–3311. 3, 4

[15] T. Sattler, B. Leibe, and L. Kobbelt, "Improving image-based localization by active correspondence search," in *European conference on computer vision*. Springer, 2012, pp. 752–765. 3

[16] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571. 3, 4

[17] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015. 3

[18] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417. 3, 4

[19] X. Yu, S. Chaturvedi, C. Feng, Y. Taguchi, T.-Y. Lee, C. Fernandes, and S. Ramalingam, "Vlase: Vehicle localization by aggregating semantic edges," *arXiv preprint arXiv:1807.02536*, 2018. 3, 5

[20] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *CVPR Deep Learning for Visual SLAM Workshop*, 2018. [Online]. Available: http://arxiv.org/abs/1712.07629 3, 4

[21] "Long-term visual localization." [Online]. Available: https://www.visuallocalization.net/workshop/eccv/2020/ 3

[22] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2. 3

[23] "Dbow3," 2017. [Online]. Available: https://github.com/rmsalinas/DBow3 4