Hindawi
Journal of Robotics
Volume 2022, Article ID 9354909, 11 pages
https://doi.org/10.1155/2022/9354909

Research Article

Learning a Robust Hybrid Descriptor for Robot Visual Localization

Qingwu Shi, Junjun Wu, Zeqin Lin, and Ningwei Qin

School of Mechatronic Engineering and Automation, Foshan University, Foshan, China

Correspondence should be addressed to Junjun Wu; jjunwu@fosu.edu.cn

Received 11 February 2022; Revised 23 March 2022; Accepted 19 April 2022; Published 19 May 2022

Academic Editor: Xianfeng Yuan

Long-term robust visual localization is one of the main challenges of long-term visual navigation for mobile robots. Due to factors such as illumination, weather, and season, mobile robots continuously navigate with visual information in a complex scene, which is likely to lead to failure localization within a few hours. However, semantic segmentation images will be more stable than the original images against considerable drastically variable environments; therefore, to make full use of the advantages of both semantic segmentation image and its original image, this paper solves the above problems with the latest work of semantic segmentation and proposes the novel hybrid descriptor for long-term visual localization, which is generated by combining a semantic image descriptor extracted from segmentation images and an image descriptor extracted from RGB images with a certain weight, and then trained by a convolutional neural network. Our experiments show that the localization performance of our method combining the advantages of semantic image descriptor and image descriptor is superior to those long-term visual localization methods with only an image descriptor or semantic image descriptor. Finally, our experimental results mostly exceed state-of-the-art 2D image-based localization methods under various challenging environmental conditions in the Extended CMU Seasons and RobotCar Seasons datasets in specific precision metrics.

1. Introduction

Visual localization is a key part of SLAM for mobile robots, which can help the robot to determine the general position and direction. In the GPS-constrained environment, it plays a vital role in navigation for mobile robots [1]. When the robot performs visual navigation, it usually generates an environmental map based on scene representation under certain environmental conditions. However, due to the influence such as weather, illumination, and season, when the robot moves in a large range for a long time, environmental conditions of the scene image that is being observed may vary greatly from the previous. Therefore, long-term visual localization methods need to deal with all these appearance variations [2, 3]. At present, various visual localization problems under the more challenging environment have attracted extensive attention of researchers [4–8]. Therefore, this paper mainly focuses on the long-term visual localization problems for the robots under complex environments, namely, finding the pose of the image in the currently constructed map which is highly similar to the currently observing image. With the above preliminary global localization coarsely, the initial pose can be provided for the regression of local high-precision 6-DOF camera pose with the hierarchical localization methods [9–11].

Traditional methods such as SIFT, SURF, and ORB rely on point descriptor for visual localization. Recently, the global image descriptor [12] extracted from the deep convolutional neural network performs better than the above traditional point descriptor. However, these methods only aggregate the features of the image region without considering the semantic information contained therein.

Intuitively, one of the main challenges of mobile robots when performing long-term work is still to obtain the representation of images under changing conditions. However, semantic information of objects in the scene that is extracted by semantic segmentation or object detection can generate the invariant representation of images under

changing conditions. For example, the semantic information of a tree will not change whether or not it is covered with snow, so the visual localization methods with semantic information have attracted the attention of researchers [4–7, 13].

In summary, to improve the accuracy of long-term visual localization for the mobile robots in complex and changing environments, a novel method of long-term visual localization based on hybrid descriptor is proposed, which is generated by combining a semantic image descriptor extracted from segmentation images and image descriptor extracted from RGB images. However, the performance of semantic segmentation based on CNN highly depends on semantic labels, which are expensive and time-consuming to obtain. Therefore, to reduce the large costs of manual labeling for semantic labels, the paper introduces 3D geometric consistency supervision for the training process of segmentation network PSPNet [14], so that the segmentation effects of the same scene under changing environmental conditions are generally consistent. Finally, this paper verifies the effectiveness of this method on Extended CMU Seasons and RobotCar Seasons datasets. The contributions of this paper are as follows:

A new method of long-term visual localization based on hybrid descriptor is proposed, which is a compact hybrid descriptor generated by concatenating a semantic image descriptor extracted from semantic segmentation images and an image descriptor extracted from RGB images with a certain weight and then trained by a convolutional neural network.

This paper introduces 3D geometric consistency supervision for the training process of segmentation network PSPNet to obtain the semantic labels of the training and testing datasets with little labor costs.

This paper shows that the localization performance of our method by combining the advantages of the semantic image descriptor and image descriptor is superior to that of long-term visual localization methods with only an image descriptor or semantic image descriptor. Besides, our method is comparable to state-of-the-art 2D image-based localization methods under various challenging environmental conditions in the Extended CMU Seasons and RobotCar Seasons datasets in specific precision metrics.

The organizational structure of this paper is as follows: Section 1 introduces the research background, as well as defines the challenging problems and the contribution of our method. Section 2 reviews the research related to the work of this paper, mainly including semantic segmentation, domain adaptation, and long-term visual localization in a changing environment. In Section 3, the network architecture and the loss function of our method are described in details. In Section 4, the experimental scheme is introduced in details, including datasets, evaluation metrics and experimental results, and the analysis of the experimental results. In Section 5, the research work of this paper is summarized and the future work prospects is given.

2. Related Works

2.1. Semantic Segmentation. Semantic segmentation is the task of assigning a category label to each pixel in the input image, which is a very important task for visual perception of mobile robots. In the early stage, researchers mainly use manually designed descriptor or probability graph models to perform semantic segmentation tasks. In recent years, deep convolutional neural networks based semantic segmentation have been proved to be superior to traditional methods. The pioneering work of Long et al. [15] shows that convolutional neural networks (CNN) originally used for classification, such as AlexNet or VGG, can be transformed into fully convolutional networks (FCN) for semantic segmentation. The following work improved the structure of its neural network based on [15], such as expanding the receptive field [16, 17], making full use of global context information [14] or fusing multiscale features [18, 19]. In addition, some work combined FCN [15], with probabilistic graphical models such as conditional random fields, as a post-processing step [16].

However, the performance of semantic segmentation based on CNN highly depends on semantic tags, which are expensive and time-consuming to obtain. In this case, many weakly supervised methods have been proposed with labels in the form of such as bounding boxes [20], image-level tags [21], or points [22]. In addition, [23] obtains semantic tags in a semiautomatic way, which requires low manual costs than pixel-level annotation to improve the segmentation performance. This paper adopts the method similar to [23] to obtain the segmentation maps of mapillary street level sequences [24].

2.2. Domain Adaptation. The training of models for deep learning requires a large amount of labeled data, but manually labeling a large amount of data is time-consuming and laborious. However, the pixel-level annotation in the source task is available; therefore, the purpose of the domain adaptation method is to learn the knowledge of the source task, so that the model can perform well in the target task. Early work includes [25, 26], which converts the features of the target domain into the source feature domain [25] or the domain-invariant feature space [26]. Some researchers focus on domain adaptation based on the CNN models [27, 28]. These methods mainly aim to make the learned model obtain domain-invariant features, either by training the network based on the adversarial loss to promote the confusion between source domain and target domain [27] or by keeping the feature distribution of source domain and target domain consistent [28]. Recently, several domain adaptation methods have been proposed for semantic segmentation tasks [29–33]. Most of them [29–31] use synthetic datasets, such as [34], which can automatically generate a large number of annotated synthetic images. The method proposed in [31] utilizes an image translation-based technique, which converts the image from the source domain into the target domain and then performs segmentation tasks. Another common approach is to train the network based on

adversarial loss, such as [32], which causes the network to fool the domain discriminator to generate roughly the same feature distribution as the generator.

Although the domain adaptation method can also obtain the semantic labels of the training and test datasets, its performance is not good, we introduce 3D geometric consistency as the supervision signal for the training process of segmentation network PSPNet to obtain the semantic labels of the training and test datasets used in this paper, so that the semantic labels for images of the same scene under different environmental conditions are generally consistent.

2.3. Long-Term Visual Localization. Because the benchmark datasets proposed in [2, 3] are challenging and the evaluation metrics of the visual localization provided by it are convincing, it has greatly promoted the research and development of long-term visual localization. At present, the methods of long-term visual localization generally include: Sequence-based image retrieval methods [35], learning-based local feature localization methods [36, 37], 3D structure-based localization methods [38–40], 2D image-based localization methods [5, 12, 41–45], and hierarchical localization methods [9–11].

The 2D image-based localization methods have great advantages in robustness and efficiency, so this paper focuses on the 2D image-based visual localization methods, which do not use any form of 3D reasoning method to calculate the pose of the query image and is usually used for place recognition or closure loop detection tasks of visual SLAM. For the image that is being observed by the robot, given a set of environmental maps with a known camera pose, the 2D image-based localization methods usually approximate the pose of the image that is the most similar visual appearance in the map to the pose of currently observing image (i.e., query image). Since 2D image-based localization methods generally perform well at coarse precision, the hierarchical localization methods use the initial pose obtained by 2D image-based localization methods to regress the high precision 6-DOF camera pose further.

VLAD [46] is a classic method for 2D image-based localization or place recognition under ideal conditions, but it has poor robustness for tasks of long-term visual localization under dramatically changing conditions. On this basis, DenseVLAD [41] uses the VLAD clustering RootSIFT descriptor that is used to match for tasks of visual localization. Subsequently, the long-term visual localization methods based on 2D image localization have achieved good development with the help of CNN models. Therefore, NetVLAD [12] integrates the traditional VLAD algorithm into CNN network structure to achieve end-to-end visual localization.

To improve the visual localization performance in complex environments, many works utilize semantic information [5, 13, 23, 44], context information [45], and depth information [5, 47] under the architecture of the convolutional neural network to learn scene descriptor with invariant environmental conditions. However, these tasks require auxiliary information that usually requires large labor costs to obtain. This paper also makes full use of semantic information as auxiliary information to overcome the impact such as illumination, weather, and season on visual localization tasks. However, to reduce the high manual labeling costs for obtaining auxiliary information, we introduce 3D geometric consistency as the supervision signal for the training process of segmentation network PSPNet to obtain the semantic labels of the training and testing datasets used in this paper.

3. Proposed Method

3.1. Network Model Structure. Figure 1 shows the network model structure of the long-term visual localization method designed in this paper. Above all, this paper adopts a method similar to [23] to obtain the segmentation images, and the specific steps are as follows: (1) using the 2D-2D matches that are composed of two images of the same scene taken under different conditions provided by [23], which provides constraints for the training process of the segmentation network PSPNet, i.e., the segmentation maps of the two images in the same scene should be consistent and (2) using the cross-season correspondence dataset in [23], some roughly annotated images and the correspondence loss, the segmentation effects of images with the same scene can be roughly consistent under changing environmental conditions. For more details, please refer to [23].

The network model structure consists of four parts: (1) training a VGG16 network to extract 16K (1-dimensional) semantic image descriptors from segmentation images, (2) training another VGG16 network to extract 16 K (1-dimensional) image descriptors from RGB images, (3) concatenating the descriptors of semantic image descriptors and image descriptors with the weight of λ_1 and λ_2 ($\lambda_1 + \lambda_2 = 1$), respectively, to obtain 16K (1-dimensional) hybrid descriptors, and (4) training a convolutional neural network that is composed of three convolutional layers and two fully connected layers, which converts 16K (1-dimensional) hybrid descriptors to 1024 (1-dimensional) learning hybrid descriptors for tasks of visual localization.

When the mobile robot builds the environment map incrementally, each image that is being observed is processed with our method to generate 1024 (1-dimensional) learning hybrid descriptors for the visual localization module. This descriptor is feature data that contain invariant representation of image, so the environment map built by the robot exists in the form of feature database. The main task of the visual localization module is to continuously measure the similar distance between the feature data that are generated from the currently observing image with our method and feature database based on L1 distance, and the pose of the currently observing image is approximated to the known pose of the candidate image with the lowest similar distance.

3.2. Loss Function. As shown in Figure 1, this paper needs to optimize two types of loss functions: total loss of segmentation task for obtaining semantic segmentation images and triplet loss for task of visual localization.

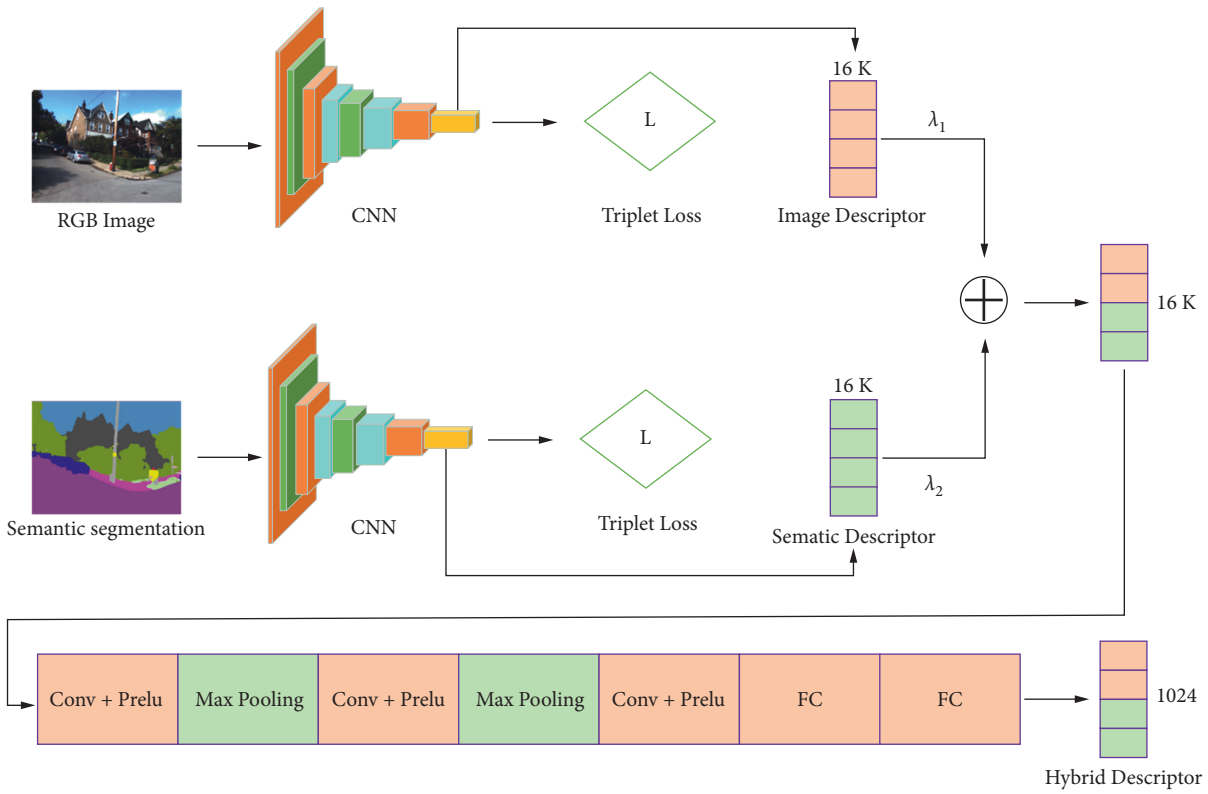

Figure 1: Overview of the network model structure.

3.3. Total Loss for Segmentation Task. In order to obtain the high-quality segmentation images used in this paper with little labor costs, we introduce 3D geometric consistency as the supervision signal for the training process of segmentation network PSPNet and need to optimize the loss function for segmentation task, which is composed of standard cross-entropy loss function $\mathscr{L}_{\text{sce}}$ and correspondence loss $\mathscr{L}_{cor}$ to obtain the high-quality semantic segmentation images used in this paper. The total loss function for segmentation task, i.e., $\mathscr{L}$ is defined as follows:

$$\mathscr{L} = \mathscr{L}_{\text{sce}} + \omega \mathscr{L}_{\text{corr}}, \tag{1}$$

where ω is the weight, we set ω to 1, and the correspondence loss $\mathscr{L}_{cor}$ is designed as follows:

$$\mathscr{L}_{\text{cor}} = \sum_{(r,t)} l(I_r, I_t, p_r, p_t), \tag{2}$$

where I_r is the reference traversal image, I_t is the target traversal image, p_r and p_t are the pixel positions of matching points in the reference traversal and target traversal images, respectively. l is hinge loss or the correspondence cross-entropy loss, for more details about l, please refer to [23].

3.4. Triplet Loss. In order to guide the model learn robust descriptor for visual localization tasks, we construct triplet loss in the process of training a VGG16 network to extract 16K (1-dimensional) semantic image descriptors from segmentation images and training another VGG16 network to extract 16K (1-dimensional) image descriptors from RGB images. In order to enhance the robustness of the descriptor used for visual localization, this paper obtains tuples from the training dataset to train the network. Tuples are composed of an anchor image (p), a positive sample (q, i.e., the same scene as the anchor image), and i negative samples (n_i, i.e., different scenes from the anchor image). To make the distance between positive pairs decrease and make the distance between negative pairs increase, the triplet loss used in this paper is as follows:

$$\mathscr{L}_T = \sum_i \max\left(0, 1 - \frac{\|p - q\|_2}{m + \|p - n_i\|_2}\right), \tag{3}$$

where m means margin, we set m to 0.1, p, q, and n_i refer to the cached embeddings for the anchor, positive, and negative images.

4. Experiment

This section describes the experimental protocol in details, including experimental dataset, experimental settings, evaluation metrics, comparison models, experimental results, visualization experiment, and ablation experiment.

4.1. Experimental Dataset

4.1.1. Training Dataset. Mapillary street level sequences [24] are currently the most diverse publicly available dataset for long-term place recognition tasks, covering the regional environment of 30 major cities across six continents from

Tokyo to San Francisco for more than seven years, which contains more than 1.6 million images collected from the mapillary and contains huge perceptual changes due to dynamic objects, seasons, region, weather, cameras, and illumination.

4.1.2. Testing Dataset. The Extended CMU Seasons dataset [2] is a subset of the CMU Visual Localization dataset [48]. It records the scene images in a variety of challenging environments such as suburban, urban, and park in Pittsburgh of the United States for more than a year. This dataset contains a total of 1 reference traversal and 11 query traversals, the environmental conditions of the images in the reference traversal are Sunny + No Foliage. The 11 query traversals contain different regional environments (suburban, urban, and Park), different vegetation conditions (no foliage, mixed foliage, and foliage) and different weather conditions (sunny, cloudy, low sun, overcast, and snow), respectively.

RobotCar Seasons dataset [2] is from the publicly available Oxford RobotCar dataset [49], which records the scene images with various changing conditions in Oxford, UK for one year. This dataset contains a total of 1 reference traversal (environmental condition is overcast) and 9 query traversals that contain 5 weather conditions (snow, dusk, sun, dawn, and rain), 2 types of seasons (overcast winter and overcast summer) and 2 night environmental conditions (night and night rain). The environmental conditions of the latter two query traversal constitute Night All, which forms a comparison of different illumination conditions with the Day All formed by the environment of the previous seven query traversals.

4.1.3. Evaluation Metrics. Reference [2] hosts a performance evaluation server for evaluating different visual localization methods, which has attracted extensive attention from many researchers. Therefore, our experiments use the metrics in [2] to test the visual localization performance of the proposed method. The experiments upload the 6-DOF pose files of the query image obtained by our method to this server, and we obtain the performance results and ranking on the public evaluation website. There are three precision metrics on this evaluation site: high precision $(0.25\,m, 2^{\circ})$, medium precision , $(0.5\,m, 5^{\circ})$, and coarse precision $(5\,m, 10^{\circ})$. The website calculates the percentage of pose error within the three precision metrics to evaluate the performance of various visual localization methods.

4.2. Comparison Models. In the experiment, four typical and advanced methods are selected as the comparison model of the method in this paper, which is shown as follows. NetVLAD [12] realized the 2D image-based localization task by integrating the classic VLAD algorithm into the CNN model structure. DenseVLAD [41] realized the 2D image-based localization task by using the VLAD clustering RootSIFT descriptor. WASABI [32] proposed a global image descriptor integrating semantic and topological information constructed by wavelet transform based on semantic edge, which realized 2D image-based localization task by matching semantic edge transformation. DIFL [29] introduced feature consistency loss to train the encoder to generate domain-invariant features in a self-supervised manner to achieve 2D image-based localization tasks.

4.3. Experimental Setting. In this paper, the mapillary street level sequence dataset and the segmentation maps obtained by the method similar to [23] are used for model training. We adopts the Extended CMU Seasons and RobotCar Seasons datasets for testing performance of our method, and the test results were uploaded to the long-term visual localization performance evaluation website provided in [2, 3] (https://www.visuallocalization.net/).

We implemented the proposed method using Pytorch on the computer with two 2080Ti GPUs. The training dataset of this experiment was resized to $640 * 480$, and we trained mapillary street level sequences using the ADAM optimizer, and batch size was set to 8 tuples (containing no more than 15 negative samples). The initial learning rate is set to 0.0002, the margin is set to 0.1, and epochs are set to 30.

4.4. Testing Experiment on the Extended CMU Season Dataset. The testing files obtained by the proposed method were uploaded to the above visual localization performance evaluation website. Several state-of-the-art 2D image-based localization methods in this website are selected to compare the visual localization performance with our method under different regional environments, vegetation conditions, and weather conditions.

4.5. Environment under Different Regional Environments. The localization performance of the proposed method and the selected comparison models under different regional environments in the Extended CMU Seasons dataset is shown in Table 1. When the mobile robots move in a large range, environmental conditions of the scene image that is being observed may be significantly different from the previous, so the long-term visual localization method needs to cover as many regional environments as possible. Therefore, the testing environments selected in the experiment include three typical regional environments, namely, urban, suburban, and park. According to the data in Table 1, for suburban and park environments, the performance of the model in this paper is 14.59% and 14.62% higher than the state-of-the-art baselines under the coarse precision metric. In the urban environment, except for the coarse precision metric where our model ranks second in performance, our model performs the best in other cases.

It can be seen that the proposed method is significantly advanced in the park and suburban environments, and its performance is weakened in the urban environments, but it is still more competitive than other state-of-the-art baselines. This is mainly because there are a large number of trees and other static objects in the park and suburban environments, and the proposed model can make full use of the semantic information of these two types of scenes to enhance

Table 1: Results of different regional environments.

Method	Park 0.25 m/0.5 m/5m 2°/5°/10°	Suburban 0.25 m/0.5 m/5m 2°/5°/10°	Urban 0.25 m/0.5 m/5m 2°/5°/10°
NetVLAD [12]	2.6/10.4/55.9	3.7/13.9/74.7	12.2/31.5/**89.8**
DIFL-FCL [29]	6.1/20.7/69.1	5.6/18.2/69.8	14.8/35.1/79.6
DenseVLAD [41]	5.2/19.1/62.0	5.3/18.7/73.9	14.7/36.3/83.9
WASABI [32]	2.4/9.1/54.5	3.8/13.9/67.3	7.9/21.3/75.2
Ours	**7.0/24.5/79.2**	**6.1/20.7/85.6**	**16.1/39.0**/87.7

Table 2: Results of different vegetation conditions.

Method	Foliage 0.25 m/0.5 m/5m 2°/5°/10°	Mixed foliage 0.25 m/0.5 m/5m 2°/5°/10°
NetVLAD [12]	6.2/18.5/74.3	5.8/17.6/71.1
DIFL-FCL [29]	8.2/22.2/69.0	9.6/26.0/74.4
DenseVLAD [41]	7.4/21.1/68.0	8.5/24.5/73.0
WASABI [32]	4.9/15.2/67.6	4.8/14.8/64.9
Ours	**9.5/26.5/81.2**	**10.5/29.4/86.7**

the accuracy of visual localization. Therefore, the localization performance of the model in this paper improves most under the coarse precision of these two regional environments. However, the semantic information of the same scene in the urban environment is changed due to the existence of a large number of dynamic objects such as pedestrians or cars, which makes the performance of the model in this paper affected.

In summary, the performance of the proposed method in different regional environments is significantly better than that of the selected representative existing methods. Therefore, the model presented in this paper plays a positive role in tasks of the long-term localization for mobile robots, especially in the park and suburban environments.

4.6. Environment under Different Vegetation Conditions. The results of the experiment for environments with different vegetation conditions in the Extended CMU Seasons dataset are shown in Table 2. For two complex vegetation conditions mixed foliage and foliage, the proposed method shows the best robustness compared with other state-of-the-art baselines in three precision metrics.

What is undoubtedly exciting is that different vegetation conditions are the most challenging environmental conditions in the Extended CMU Seasons dataset due to the types, numbers, and positions of leaves. The outstanding visual localization performance of the proposed method is mainly because of the segmentation images used in this paper, which makes the extracted environmental features and constructed scene descriptor have stronger invariance representation, which has a practical value for mobile robots to perform long-term outdoor navigation.

4.7. Environment under Different Weather Conditions. Mobile robots are inevitably faced with weather variations when they work for a long time. Therefore, we not only tested the effectiveness under different regional environments and different vegetation conditions in the Extended CMU Seasons dataset but also tested it in different weather conditions. The experimental results are shown in Table 3. It can be seen from Table 3 that our model outperforms other state-of-the-art baselines with most weather conditions, and our method performs even more prominently under the overcast and low sun environments in the three precision metrics especially.

4.8. Testing Experiment on RobotCar Seasons Dataset. The experiment in this section uses the same method as that in Section 4.5 to verify the effectiveness of our method, and we select several state-of-the-art 2D image-based localization methods in the evaluation website for performance comparison. The experimental environments include two kinds of changing conditions: different weather and illumination conditions.

4.9. Testing under Different Weather Conditions. The comparison results of testing the robustness between the model proposed in this paper and three existing comparison models under different weather conditions in the RobotCar Seasons dataset are shown in Table 4. The proposed method has the best localization performance in the medium and high precision metrics under the snow condition. In addition, although there are a large number of dynamic targets such as pedestrians or cars in the RobotCar Seasons dataset, which changes the semantic information in the same scene, the model in this paper still achieves decent results.

4.10. Testing under Different Illumination Conditions. In addition, we also conducted experiments under different illumination conditions. The test results are shown in Table 5. Compared with the other three comparison models, the proposed model improved 24.00% and 24.62%, respectively, in terms of high and medium precision metrics under night conditions, which is undoubtedly exciting because night conditions are the most challenging environmental conditions in the RobotCar Seasons dataset.

4.11. Visualization Experiment. For the Extended CMU Seasons and RobotCar Seasons datasets, we obtained segmentation images using self-supervised methods similar to [23], as shown in Figures 2 and 3, respectively. Because the ultimate purpose of this paper is to take the obtained segmentation images as input and train the model to generate semantic image descriptor to improve the performance of visual localization tasks, the predicted images of semantic segmentation in the same scene in different environments under the Extended CMU Seasons and RobotCar Seasons datasets should be consistent. As can be seen from Figures 2 and 3, the predicted image of semantic segmentation in the same scene under different environments in the extended CMU seasons and RobotCar Seasons datasets are generally the same.

Table 3: Results of different weather conditions.

Method	Overcast 0.25 m/0.5 m/5m 2°/5°/10°	Low sun 0.25 m/0.5 m/5m 2°/5°/10°	Cloudy 0.25 m/0.5 m/5m 2°/5°/10°	Snow 0.25 m/0.5 m/5m 2°/5°/10°
NetVLAD [12]	6.7/19.1/76.3	5.5/17.5/71.3	6.8/20.1/79.5	5.0/16.4/68.0
DIFL-FCL [29]	9.7/25.3/70.9	8.7/25.3/74.4	8.8/24.7/76.9	7.4/26.7/73.5
DenseVLAD [41]	8.4/23.3/72.1	8.3/26.1/76.0	**9.3**/**27.3**/80.5	**8.3**/**29.0**/78.9
WASABI [32]	5.4/15.8/70.8	4.2/14.0/62.1	5.1/15.3/71.0	3.4/13.2/58.0
Ours	**11.0**/**29.3**/**84.4**	**9.2**/**28.0**/**85.5**	**9.3**/27.1/**88.2**	7.0/24.5/**79.2**

Table 4: Results of different weather conditions.

Method	Dawn 0.25 m/0.5 m 2°/5°	Sun 0.25 m/0.5 m 2°/5°	Snow 0.25 m/0.5 m 2°/5°
NetVLAD [12]	6.2/22.8	5.7/16.5	7.0/25.2
DIFL-FCL [29]	9.5/30.2	**9.1**/23.7	9.0/25.2
DenseVLAD [41]	8.7/**36.9**	5.7/16.3	8.6/30.1
Ours	**10.1**/35.4	8.9/**25.9**	**10.8**/**30.3**

Table 5: Results of different illumination conditions.

Method	Night all 0.25 m/0.5 m 2°/5°	Day all 0.25 m/0.5 m 2°/5°
NetVLAD [12]	0.3/2.3	6.4/26.3
DIFL-FCL [29]	2.5/6.5	7.6/26.2
DenseVLAD [41]	1.0/4.4	7.6/31.2
Ours	**3.3**/**8.1**	**8.6**/30.5

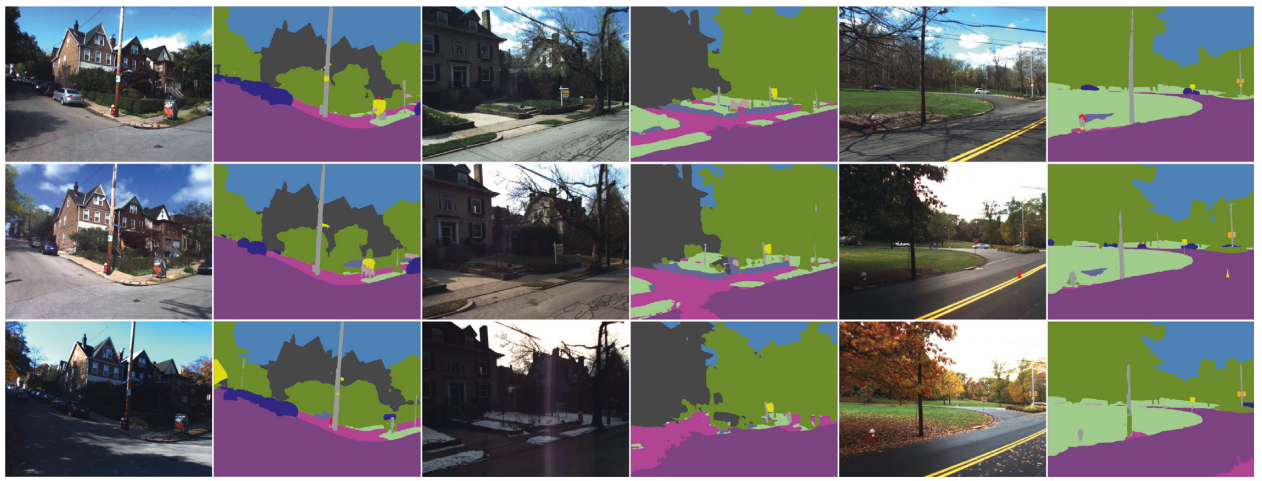

Figure 2: Predicted images of semantic segmentation in the same scene under different environments in the Extended CMU Seasons dataset.

4.12. Experiment for Testing the Proposed Algorithm. Considering the computational efficiency of the proposed algorithm, we also test the average computational time for retrieval in different database images. This paper focuses on the time when the query image matches the database image. RobotCar Seasons and Extended CMU Seasons datasets are used in the experiment, and the total database images of both are 20,862 and 10,338, respectively. As can be seen from Table 6, the algorithm queries each frame taking 31.42 ms in the Extended CMU Seasons dataset (10k database images), that is, the algorithm can process about 32 frames per second on average, while the algorithm takes 52.17 ms to query each frame in the RobotCar Seasons dataset (20k database images), that is, the algorithm can process about 19 frames per second on average. However, usually the acquisition frequency of mobile robots is 15 frames per second, and a large number of invalid frames need to be removed, which shows that our algorithm has a practical value for mobile robots.

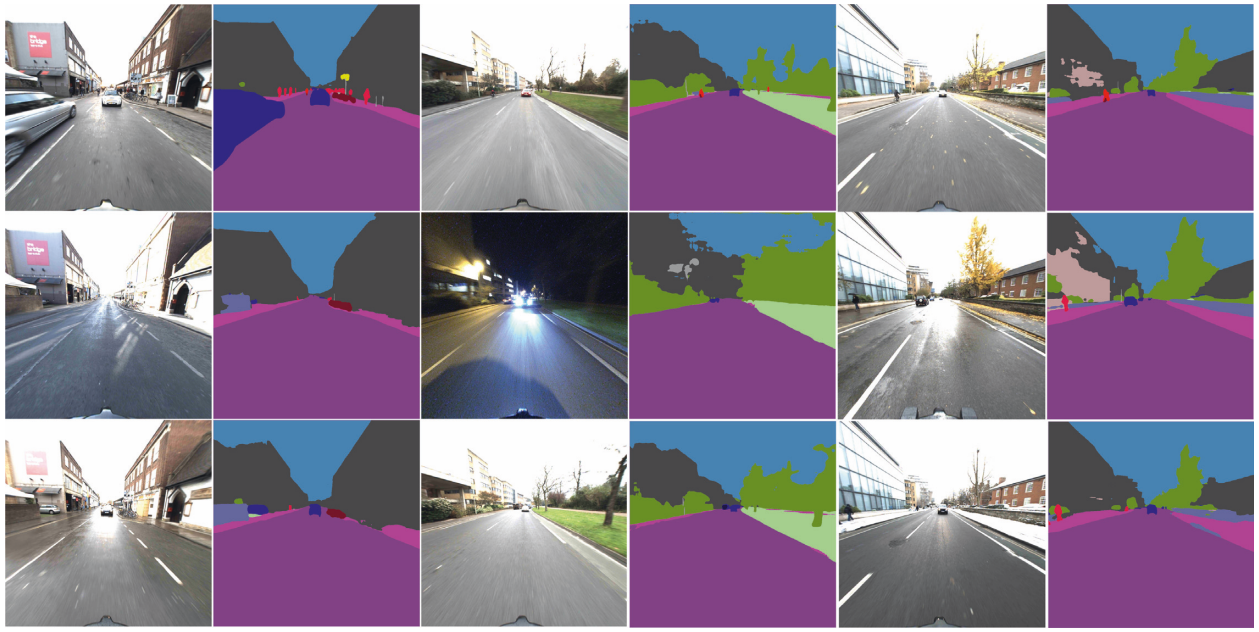

Figure 3: Predicted images of semantic segmentation in the same scene under different environments in the RobotCar Seasons dataset.

Table 6: Query time per frame of database images in different datasets.

Dataset	RobotCar seasons (20k)	Extended CMU seasons (10 K)
Query time (ms)	52.17	31.42

Table 7: Ablation study for our method with different weights λ_1 and λ_2.

Training/ Testing		Park	Suburban	Urban	Night all	Day all
λ_1	λ_2	0.25 m/0.5 m/5m 2°/5°/10°	0.25 m/0.5 m/5m 2°/5°/10°	0.25 m/0.5 m/5m 2°/5°/10°	0.25 m/0.5 m/5m 2°/5°/10°	0.25 m/0.5 m/5m 2°/5°/10°
0	1	2.1/8.3/49.2	2.9/10.9/65.4	11.9/28.9/73.9	1.1/2.8/11.3	4.7/15.4/54.7
0.1	0.9	3.6/10.1/52.4	3.9/13.7/74.1	12.5/32.6/77.8	1.9/6.1/14.5	5.4/20.1/63.0
0.2	0.8	4.9/20.1/75.3	5.1/16.7/75.3	13.1/34.4/78.5	2.2/6.7/18.7	6.1/22.3/71.0
0.3	0.7	5.5/21.3/76.1	5.4/20.4/77.3	14.6/35.7/79.2	1.9/7.3/21.1	5.8/22.8/73.2
0.4	0.6	6.2/21.2/76.9	5.3/19.7/77.5	15.6/37.3/80.8	2.4/6.4/18.1	6.5/21.3/74.3
0.5	0.5	6.5/22.1/79.1	5.8/21.3/76.9	15.7/37.6/81.2	2.7/6.7/17.3	6.3/23.4/73.2
0.6	0.4	6.4/23.3/77.5	5.5/20.1/78.6	15.7/37.5/81.0	2.4/7.8/19.7	7.2/24.1/72.1
0.7	0.3	6.9/24.1/78.7	5.9/20.4/79.6	15.9/37.9/81.4	3.0/7.7/22.1	8.2/29.1/76.5
0.8	0.2	6.7/23.9/79.1	5.7/20.9/85.0	16.0/38.7/82.0	3.3/8.1/26.1	8.6/30.5/82.2
0.9	0.1	7.0/24.5/79.2	6.1/20.7/85.6	16.1/39.0/87.7	2.6/6.9/18.9	6.8/25.3/71.0
1	0	5.6/11.2/69.1	4.9/17.2/79.3	14.1/35.4/83.5	1.7/3.5/14.5	5.7/19.6/65.3

4.13. Ablation Study. A VGG16 network is trained from semantic segmentation images to extract a 16K (1-dimensional) semantic image descriptor and another VGG16 network is trained from RGB images to extract the 16K (1-dimensional) image descriptor to concatenate with weights λ_1 and λ_2 ($\lambda_1 + \lambda_2 = 1$), respectively. Therefore, this section discusses the influence of different weights λ_1 and λ_2 on the performance of the proposed method. The experimental results are shown in Table 7.

$\lambda_1 = 0$, $\lambda_2 = 1$ means that only one VGG16 network is trained from semantic segmentation images to extract the 16K (1-dimensional) semantic image descriptor, while $\lambda_1 = 1$, $\lambda_2 = 0$ means that only one VGG16 network is trained from RGB images to extract the 16K (1-dimensional) image descriptors directly for visual localization tasks. Park, suburban, and urban in Table 7 are the regional environmental conditions in the Extended CMU Seasons dataset, while Day All and Night All in Table 7 are the illumination variation conditions in the RobotCar Seasons dataset. As can be seen from Table 7, the visual localization performance is the best when $\lambda_1 = 0.9$, $\lambda_2 = 0.1$ for the regional environmental conditions in the Extended CMU Seasons dataset, while the visual localization performance is the best when $\lambda_1 = 0.8, \lambda_2 = 0.2$ for the illumination variation conditions in

the RobotCar Seasons dataset, which shows that if we trust image descriptor more than semantic image descriptor, we can reach the best results for both datasets, and the semantic image descriptor is more helpful for the illumination condition in the RobotCar Seasons dataset compared with the regional environment in the Extended CMU Seasons dataset. Besides, it can be seen from Table 7 that for the regional environmental conditions in the Extended CMU Seasons dataset and the illumination variation conditions in the RobotCar Seasons dataset, the visual localization effect is not ideal if solely using the semantic image descriptor that is trained from semantic segmentation images or solely using the image descriptor that is trained from RGB images.

5. Conclusion

Aiming at the challenges of robustness faced by mobile robot when it performs long-term work under complex changing conditions, a new method of long-term visual localization based on hybrid descriptor is proposed, which is a compact hybrid descriptor generated by concatenating a semantic image descriptor extracted from semantic segmentation images and the image descriptor extracted from RGB images with a certain weight, and then trained by a convolutional neural network. In this paper, we verify that the visual localization performance of solely using q semantic image descriptor trained from semantically segmented images or solely using image descriptor trained from RGB images is not better than that of using hybrid descriptor obtained by the combination of both with a certain weight. This model was trained on mapillary street level sequences dataset and subsequently tested on Extended CMU Seasons and RobotCar Seasons datasets. The experimental results verify that the visual localization performance of the proposed method is significantly better than that of other state-of-the-art baselines in the Extended CMU Seasons and RobotCar Seasons datasets under different regions, vegetation conditions, weather, and illumination conditions. It can meet the requirements for mobile robots to perform long-term visual localization tasks in a variety of complex environments.

The performance of the visual localization method in this paper depends on the performance of the semantic segmentation method we choose. In addition, the depth information of the object in the same scene is proved to still have strong stability under changing environmental conditions. Therefore, we will integrate the depth information to process the visual variation between images and explore the impact of different semantic segmentation methods on the performance of the proposed method in the future.

Data Availability

All data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Key Area Research Projects of Universities of Guangdong Province under Grant 2019KZDZX1026, in part by the Natural Science Foundation of Guangdong Province under Grant :501100003453, in part by the Innovation Team Project of Universities of Guangdong Province under Grant 2020KCXTD015, in part by Free Exploration Foundation of Foshan University under Grant 2020ZYTS11.

References

[1] D. Li, "Dxslam: a robust and efficient visual slam system with deep features," in *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4958–4965, Las Vegas, LV, USA, January 2020.

[2] T. Sattler, "Benchmarking 6dof outdoor visual localization in changing conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8601–8610, Salt Lake City, UT, USA, June 2018.

[3] C. Toft, "Long-term visual localization revisited," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2074–2088, 2022.

[4] Y. You, "MISD-SLAM: multimodal semantic SLAM for dynamic environments," in *Proceedings of the Wireless Communications and Mobile Computing 2022*, Dubrovnik, Croatia, June 2022.

[5] J. Wu, Q. Shi, Q. Lu, X. Liu, X. Zhu, and Z. Lin, "Learning invariant semantic representation for long-term robust visual localization," *Engineering Applications of Artificial Intelligence*, vol. 111, Article ID 104793, 2022.

[6] J. Ni, "An improved deep residual network-based semantic simultaneous localization and mapping method for monocular vision robot," *Computational Intelligence And Neuroscience 2020*, vol. 2020, Article ID 7490840, 14 pages, 2020.

[7] J. Li, "Loop closure detection based on image semantic segmentation in indoor environment," *Mathematical Problems in Engineering*, vol. 2022, Article ID 7765479, 14 pages, 2022.

[8] M. Aladem, S. Baek, and S. A. Rawashdeh, "Evaluation of image enhancement techniques for vision-based navigation under low illumination," *Journal of Robotics*, vol. 2019, Article ID 5015741, 15 pages, 2019.

[9] P.-E. Sarlin, "From coarse to fine: robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12716–12725, Long Beach, CA, USA, June 2019.

[10] H. Germain, G. Bourmaud, and V. Lepetit, "Sparse-to-dense hypercolumn matching for long-term visual localization," in *Proceedings of the 2019 International Conference on 3D Vision (3DV)*, pp. 513–523, Québec City, QC, Canada, September 2019.

[11] T. Shi, "Visual localization using sparse semantic 3D map," in *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, pp. 315–319, Taipei, China, September 2019.

[12] R. Arandjelovic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5297–5307, Las Vegas, NV, USA, June 2016.

[13] M. Larsson, "Fine-grained segmentation networks: self-supervised segmentation for improved long-term visual localization," in *Proceedings of the IEEE/CVF International*

Conference on Computer Vision, pp. 31–41, Seoul, South Korea, October 2019.

[14] H. Zhao, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, Honolulu, HI, USA, July 2017.

[15] J. Long, E. Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, Boston, MA, USA, June 2015.

[16] L.-C. Chen and G. I. K. A. L. Papandreou, "Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[17] Yu Fisher and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, https://arxiv.org/abs/1511.07122.

[18] L.-C. Chen et al., "Attention to scale: scale-aware semantic image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3640–3649, Las Vegas, NV, USA, June 2016.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, Munich, Germany, October 2015.

[20] K. Anna, "Simple does it: weakly supervised instance and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 876–885, Honolulu, HI, USA, July 2017.

[21] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5688–5696, Venice, Italy, October 2017.

[22] B. Amy, "What's the point: semantic segmentation with point supervision," in *Proceedings of the European Conference on Computer Vision*, pp. 549–565, Springer, Amsterdam, The Netherlands, October 2016.

[23] M. Larsson, "A cross-season correspondence dataset for robust semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9532–9542, Long Beach, CA, USA, June 2019.

[24] F. Warburg, "Mapillary street-level sequences: a dataset for lifelong place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2626–2635, Seattle, WA, USA, June 2020.

[25] B. Kulis, K. Saenko, and Trevor Darrell, "What you saw is not what you get: domain adaptation using asymmetric kernel transforms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR 2011*, pp. 1785–1792, Colorado Springs, CO, USA, June 2011.

[26] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proceedings of the European Conference on Computer Vision*, pp. 213–226, Springer, Crete, Greece, September 2010.

[27] E. Tzeng, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, Honolulu, HI, USA, July 2017.

[28] M. Long, "Unsupervised domain adaptation with residual transfer networks," 2016, https://arxiv.org/abs/1602.04433.

[29] Y. Chen, Li Wen, and L. Van Gool, "Road: reality oriented adaptation for semantic segmentation of urban scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7892–7901, Salt Lake City, UT, USA, June 2018.

[30] Yi-H. Tsai, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7472–7481, Salt Lake City, UT, USA, June 2018.

[31] S. Sankaranarayanan, "Learning from synthetic data: addressing domain shift for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3752–3761, Salt Lake City, UT, USA, June 2018.

[32] M. Wulfmeier, A. Bewley, and I. Posner, "Incremental adversarial domain adaptation for continually changing environments," in *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4489–4495, IEEE, Brisbane, Australia, May 2018.

[33] X. Wu, "DANNet: a one-stage domain adaptation network for unsupervised nighttime semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15769–15778, Nashville, TN, USA, June 2021.

[34] G. Ros, "The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3234–3243, Las Vegas, NV, USA, June 2016.

[35] T. Naseer, "Robust visual robot localization across seasons using network flows," in *Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence*, Quebac, Canada, July 2014.

[36] R. Clark, "Vidloc: a deep spatio-temporal model for 6-dof video-clip relocalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6856–6864, Honolulu, HI, USA, July 2017.

[37] Z. Chen, "Deep learning features at scale for visual place recognition," in *Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3223–3230, IEEE, Marina Bay Sands, Singapore, June 2017.

[38] L. Liu, H. Li, and Y. Dai, "Efficient global 2d-3d matching for camera localization in a large-scale 3d map," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2372–2381, Venice, Italy, October 2017.

[39] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1744–1756, 2017.

[40] L. Svärm and O. F. M. Enqvist, "City-scale localization for cameras with known vertical direction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 7, pp. 1455–1461, 2017.

[41] A. Torii, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817, Boston, MA, USA, June 2015.

[42] H. Hu, "Retrieval-based localization based on domain-invariant feature learning under changing environments," in *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3684–3689, IEEE, Macau, China, November 2019.

[43] A. Benbihi, "Image-based place recognition on bucolic environment across seasons from semantic edge description," in *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3032–3038, IEEE, Paris, France, August 2020.

[44] H. Hu, Z. Qiao, M. Cheng, Z. Liu, and H. Wang, "DASGIL: domain adaptation for semantic and geometric-aware image-based localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 1342–1353, 2021.

[45] Z. Xin, "Localizing discriminative visual landmarks for place recognition," in *Proceedings of the 2019 International Conference on Robotics and Automation (ICRA)*, pp. 5979–5985, IEEE, Montréal, Canada, May 2019.

[46] H. Jégou, "Aggregating local descriptors into a compact image representation," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3304–3311, San Francisco, CA, USA, June 2010.

[47] N. Piasco and D. V. C. Sidibé, "Improving image description with auxiliary modality for visual localization in challenging conditions," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 185–202, 2021.

[48] H. Badino, D. Huber, and T. Kanade, "Visual topometric localization," in *Proceedings of the 2011 IEEE Intelligent Vehicles Symposium (IV)*, pp. 794–799, IEEE, Baden, Germany, July 2011.

[49] W. Maddern and G. C. P. Pascoe, "1 year, 1000 km: the oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.