applied sciences

MDPI

Article

Hierarchical Visual Place Recognition Based on Semantic-Aggregation

Baifan Chen [1], Xiaoting Song [1,*], Hongyu Shen [1] and Tao Lu [2,*]

[1] School of Automation, Central South University, Changsha 410083, China; chenbaifan@csu.edu.cn (B.C.); 204612194@csu.edu.cn (H.S.)
[2] Hubei Key Laboratory of Intelligent Robot, School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China
* Correspondence: me_xiaoting@csu.edu.cn (X.S.); lut@wit.edu.cn (T.L.); Tel.: +86-173-0748-2413 (X.S.)

Abstract: A major challenge in place recognition is to be robust against viewpoint changes and appearance changes caused by self and environmental variations. Humans achieve this by recognizing objects and their relationships in the scene under different conditions. Inspired by this, we propose a hierarchical visual place recognition pipeline based on semantic-aggregation and scene understanding for the images. The pipeline contains coarse matching and fine matching. Semantic-aggregation happens in residual aggregation of visual information and semantic information in coarse matching, and semantic association of semantic edges in fine matching. Through the above two processes, we realized a robust coarse-to-fine pipeline of visual place recognition across viewpoint and condition variations. Experimental results on the benchmark datasets show that our method performs better than several state-of-the-art methods, improving the robustness against severe viewpoint changes and appearance changes while maintaining good matching-time performance. Moreover, we prove that it is possible for a computer to realize place recognition based on scene understanding.

Keywords: hierarchical place recognition; semantic aggregation; semantic edges

Citation: Chen, B.; Song, X.; Shen, H.; Lu, T. Hierarchical Visual Place Recognition Based on Semantic-Aggregation. *Appl. Sci.* **2021**, *11*, 9540. https://doi.org/10.3390/app11209540

Academic Editor: Francesco Bianconi

Received: 5 September 2021
Accepted: 30 September 2021
Published: 14 October 2021

1. Introduction

Visual place recognition (VPR) is a core task of localization [1–3] and loop closure detection [4,5] for mobile robots, which means that robots can accurately identify the same place according to the images under different conditions [6–8].

However, VPR is a challenging problem, because it suffers from the influences of complex and time-varying environment and the factors of mobile robots. These problems occur due to some specific reasons: (1) high frequency environmental variability such as weather, light, and time of day; (2) long-term and slower environmental changes such as seasons and vegetation growth; (3) dynamic obstacles such as pedestrians, and vehicles; (4) static objects such as buildings that will also change due to engineering construction; and (5) the different orientation of the camera installed on the mobile robot and the movement of the robot. The problems above will cause viewpoint and appearance variations, meaning that there will be a lot of non-overlapping content in the image, making VPR more difficult.

To solve these problems in VPR, some researchers use the hand-craft features extracted from the images, such as Fab-Map [4]. It encodes image local features like SIFT [9] or SURFn [10] into the bag-of-words models [5] to represent the image in the form of word vectors, and realize place matching by calculating the distance between the corresponding word vectors of two images. However, the image local features are sensitive to illumination, weather, and other features. Nicosevici et al. [11] proposed a visual vocabulary-based loop-closure method, where the visual vocabularies could be built online, enabling the bag-of-words model to adapt to the dynamically changing environments. Milford et al. [12] proposed a method, combining intolerant but fast low resolution whole image matching with highly tolerant, sub-image patch matching processes to improve the accuracy of place recognition. Amato et al. [13] proposed a new image feature representation, called VLAD,

which realized image retrieval on large-scale datasets by aggregating the residuals of SIFT features in images. However, in general, the performance of traditional place recognition algorithms needs to be improved and they often fail to deal with severe viewpoint changes. However, the hand-craft features are very sensitive to illumination and weather. When the appearance of the environment changes significantly, it is difficult for the algorithm to achieve good results [14].

With the development of deep learning, the methods based on deep convolutional features outperform traditional handcraft features in many tasks in the field of computer vision. Features extracted through convolutional neural network (CNN) are deeper and more abstract, thus being non-sensitive to environmental conditions and appearance variations [15–17]. Chen et al. [18] applied the image features extracted by CNN to VPR, verifying the effectiveness of convolutional neural network in place recognition. Sunderhauf et al. [19] extracted the image features of different convolutional layers with a pre-trained AlexNet [18], so as to evaluate the robustness of that for viewpoint-variance and condition-variance, which provides a reference for the selection of convolutional features. Arandjelovic et al. improved the traditional method VLAD [13] and proposed NetVLAD [20], which replaced the traditional local features with CNN features and improved the performance. Chen et al. [21] proposed a CNN-based feature encoding method to create image representations by mining the salient patterns of images, tacking variations both in viewpoints and conditions. Although VPR methods based on CNN perform much better than traditional methods, there are few works focusing on utilizing visual semantic information, lacking a high-level understanding of the image.

Humans identify whether the place has been visited through analyzing the objects and the relationships between objects in the scene. In computer image processing, image semantic segmentation is an effective means for a computer to understand the contents of images. In recent years, image semantic segmentation has received significant attention and shown high performance in image scene understanding [22–28]. Some researchers have integrated visual semantic information into place recognition. Sourav et al. [29] proposed to use the semantics-aware higher-order layers of deep neural networks for identifying specific places under 180 degrees viewpoint reversed. They developed a descriptor normalization schemes to improve the robustness against appearance change. In subsequent studies [30], they integrated the previous work to solve three challenges in place recognition: reverse viewpoint, lateral perspective shift, and extreme appearance change. Aiming at the bucolic environments such as natural scenes with low texture and little semantic contents, but obvious appearance changes, Benbihi et al. [31] proposed a global descriptor based on image topological and semantic information to achieve place recognition by matching semantic edges between two images. These works have shown that it is possible and efficient to apply image semantic information to VPR.

Maohai et al. [32] studied a strong robust hierarchical localization method, and realized a coarse-to-fine hierarchical localization and autonomous navigation system for mobile robot based on pure vision. Emilio et al. [33] proposed an appearance-based method for topological mapping based on hierarchical decomposition of environment, and proved that the hierarchical method could reduce search space in identifying place while improve mapping accuracy in creating a map. Stephen and Milford [34] developed a new stacked hierarchical localization framework, which concatenated localization hypotheses from techniques with complementary characteristics at each layer, performing well on two challenging datasets. These works have proved that hierarchical strategy is useful and effective in reducing search space and localization.

Motivated by the works above, this paper believes that achieving efficient image-understanding-based VPR across appearance and viewpoint variations requires semantic understanding of the environment. Moreover, hierarchical strategy contributes to maintaining computational efficiency. We combine visual semantics and hierarchy and propose hierarchical place recognition based on semantic aggregation, to minimize the influences of appearance and viewpoint variations.

2. Hierarchical Visual Place Recognition Based on Semantic-Aggregation

Research [19] shows that the features extracted from the middle layer of CNN exhibit strong robustness against the severe image appearance changes caused by illumination, season, or weather conditions. On the contrary, high-level features are more semantically meaningful and more robust with respect to viewpoint variations.

We propose a novel coarse-to-fine hierarchical method based on semantic aggregation, making use of the mid-level convolutional features and semantic features to realize place recognition. Figure 1 shows the whole process. Our approach is a coarse-to-fine visual place recognition pipeline, and contains two parts: coarse matching and fine matching.

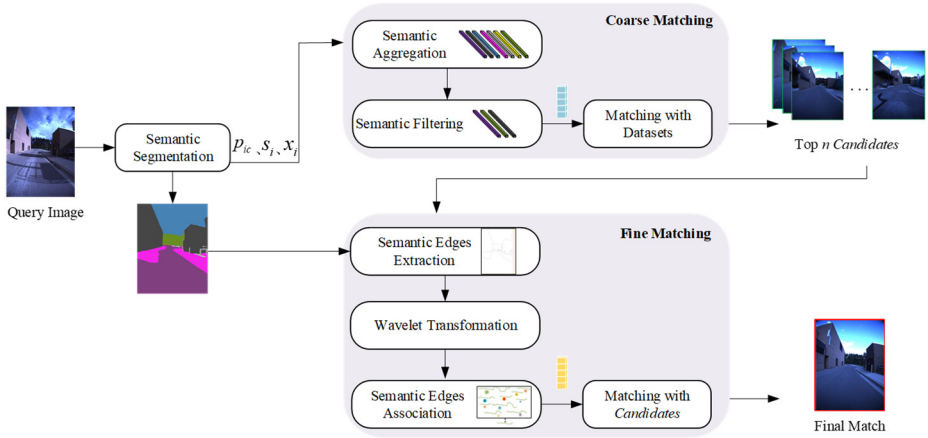


Figure 1. Overview of our proposed method. (1) Coarse matching: we get the top *n Candidates* with a global dataset search based on semantic aggregation and semantic filtering. (2) Fine matching: we select the best match through semantic edges and semantic association in the *Candidates*. Coarse matching helps to locate the query quickly, and fine matching helps to match the query accurately. Such a coarse-to-fine hierarchical progress improves the accuracy of place recognition and maintain computational efficiency.

2.1. Coarse Matching

We propose a simple yet efficient way of image representation, a hybrid global image descriptor, which can be obtained by aggregating semantic residuals for each semantic labels and semantic labels filtering. Then, we match the query with reference datasets by calculating cosine distance, to find the images with top-n similarity in the query. Those images contribute to the *Candidates*.

The whole process of coarse matching is illustrated in Figure 2. We use an advanced cross-season semantic segmentation model [35] to obtain semantic labels and their probabilities, image features, and image segmentation. This model is based on the PSP-Net [27] and it greatly improves the robustness to seasonal changes by adding enforcing label consistency across matching.

Firstly, mid-level convolutional feature map with the size of $W \times H \times D$ is extracted from the pre-trained ResNet [36] model with the dilated network strategy [37,38], where W, H, and D are the width, height, and depth of the feature map, respectively. In this task, W and H are $1/8$ of the input image size, and D is 2048. Then, a pyramid pooling module is applied to gather context information and mine rich semantic information. In the pyramid module, 4-level pyramid are fused as the global prior and are concatenated with the original feature map to generate a final feature map with the size of $W \times H \times 4096$, where W and H are $1/8$ of the input image size too.

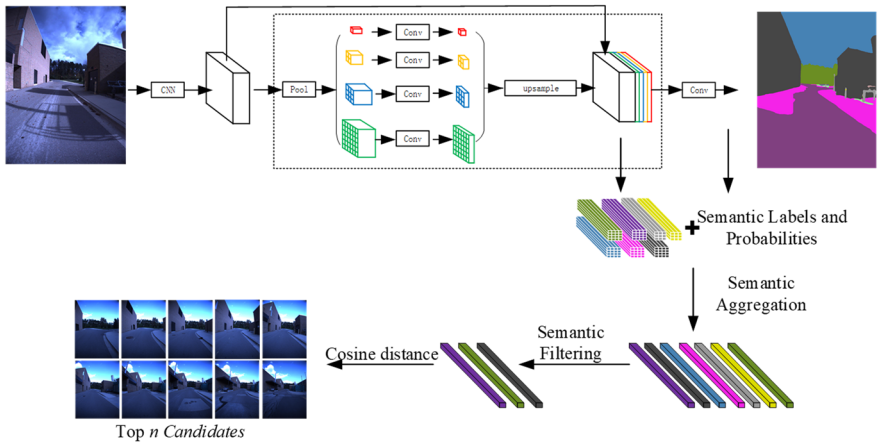


Figure 2. The process of coarse matching. Given an image, we obtain its semantic segmentation, semantic labels and probabilities, and feature maps through a network model. Then, we compute feature residuals of each semantic class and aggregate all feature residuals to get a hybrid image descriptor H_c. Subsequently, we keep the main semantic classes through semantic filtering to construct the final hybrid image descriptor H. Finally, a query is matched with the reference images by cosine distance, getting the top *n Candidates*.

Defining the semantic label s_i at position i within the feature map is as follows:

$$s_i = \underset{c}{\operatorname{argmax}}\ p_{ic}\ \ c = \{0, 1, \ldots, C\} \tag{1}$$

where c refers to the semantic classes corresponding to the related dataset, and C is the total number of semantic classes; p_{ic} represents the probability of the pixel at the location i belonging to a semantic class c.

Since each pixel's semantic class is determined, the mean descriptor m_c for each semantic class c can be computed as follow:

$$m_c = \frac{\sum_i^M \{x_i | s_i = c\}}{\sum_i^M \{i | s_i = c\}} \tag{2}$$

where x_i is the D-dimensional descriptor for the feature map, and M is the number of the pixels. Then, feature residuals of each semantic class can be computed by $|x_i - m_c|$, which preserves the distribution differences between local features and semantic mean value.

Then, we aggregate all the feature residuals of each semantic class for all the pixels in the image and weight with the corresponding semantic label probability to get H_c:

$$H_c = \sum_i^M p_{ic} |x_i - m_c| \tag{3}$$

where H_c is essentially a hybrid image descriptor based on semantic aggregation for a semantic class c.

However, using H_c for the coarse matching directly will reduce computation efficiency. Moreover, some semantic classes will reduce robustness, such as person, car, since they are dynamic. Those semantic classes will increase non-overlap contents between the images, thus leading to a low accuracy. We keep $L (L < C)$ main semantic classes to construct the final hybrid image descriptor H. H is the add of L2-normalized $H_1, H_2, \ldots, H_L$.

$$H = \left\langle \dot{H}_1 + \dot{H}_2 + \ldots + \dot{H}_L \right\rangle \tag{4}$$

where $\dot{H_1} + \dot{H_2} + \ldots + \dot{H_L}$ refer to L2-normalized $H_1, H_2, \ldots, H_L$ respectively. Meanwhile, in order to improve the ability to distinguish distance, H is normalized again as follows:

$$H = \frac{(H - m)}{\sigma} \tag{5}$$

where m and σ are the mean and standard deviation of descriptor calculated on the dataset, respectively.

After that, the query is matched with the reference images by cosine distance d_{jk}:

$$d_{jk} = 1 - \frac{H_j \cdot H_k}{\|H_j\|_2 \|H_k\|_2}, \forall j \in [1, N] \tag{6}$$

where d_{jk} is the cosine distance between the query k and reference image j in the reference dataset, and N is the number of images in reference datasets. The top n reference images with the lowest distance to the query are kept as *Candidates* and passed to the fine matching for the final match.

2.2. Fine Matching

Matching query with reference datasets only by coarse matching took a long time, imposed great pressure on the computer, and returned a low accuracy. To address this, we added a fine matching after the coarse matching to improve the matching accuracy and increase the computational efficiency, which is shown in Figure 3. A semantic edge descriptor is introduced, which does not involve the neural network calculation, and the whole process speed is fast while maintaining a high accuracy.

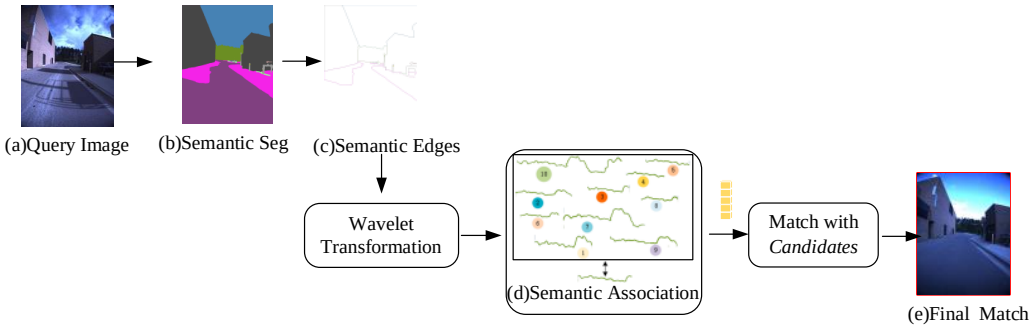


Figure 3. Illustration of fine matching process. Given an image (**a**), we first get its semantic segmentation (**b**) through coarse matching. Then, we extract its semantic edges, and describe these semantic edges (**c**) with wavelet transform. After that, we associate the query semantic edge descriptor and the Candidates semantic edge descriptors with semantic labels (**d**). Finally, we match them with cosine distance to find the best match (**e**).

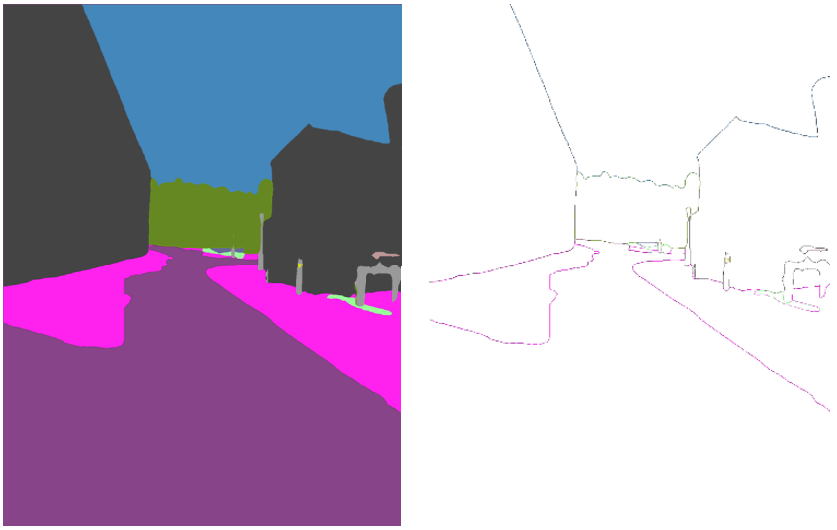
2.2.1. Semantic Edges Extraction and Description

Given an image, we can obtain its semantic segmentation through coarse matching. We firstly detect and extract its edges based on Canny, outputting a list of semantic edges and corresponding semantic labels. Figure 4 shows the semantic segmentation and its semantic edges.

There are many existing methods to describe edges. Among the existing edge descriptors, we prefer the wavelet descriptor [39]. Wavelet transform can generate a unique representation for a signal. More importantly, the multi-scale decomposition of wavelet descriptors makes the edge descriptors more compact and better discriminative.

For the semantic edges extracted above, we subsampled them and collected P pixels. Their (x, y) locations in the image are connected into a two-dimensional vector to outputting an original sequence. Then, we separately computed the discrete Harr-wavelet transform

over each row and column and normalized them by L2-normalization, outputting a semantic wavelet descriptor, which has translation, scaling, and rotation invariance.



(**a**) (**b**)

Figure 4. Semantic segmentation and corresponding semantic edges: (**a**) Semantic segmentation ; (**b**) Semantic edges.

2.2.2. Semantic Association and Matching

To make the matching more precisely, we introduced a semantic association strategy. The semantic wavelet descriptor of the query is associated with that of the reference datasets according to their semantic labels. Figure 5 shows the process of semantic edges association.

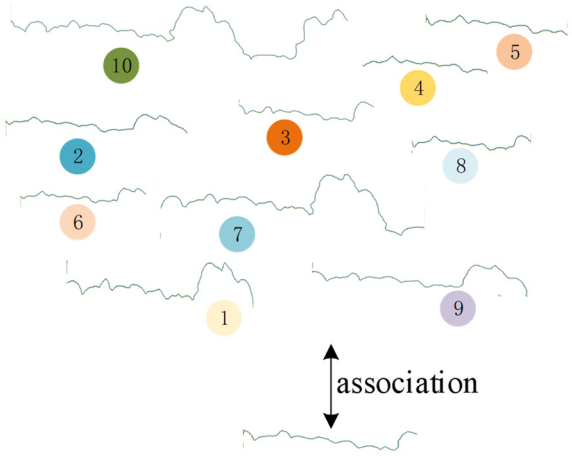


Figure 5. Semantic edges association. Edges belonging to the same semantic class are associated together to match faster and more precisely. For example, edges marked 1 to 10 are all edges labelled with vegetation in the Candidates. And edge below is the vegetation edge in the query. They are associated as a group.

3. Experiments and Results

3.1. Datasets and Performance Evaluations

We used two publicly available VPR benchmark datasets: North Campus Dataset [40] and Nordland Dataset [41], to validate the effectiveness of our method. These two datasets include viewpoint variations and appearance variations caused by seasonal changes, collection tools, and so on. Their key information is summarized in Table 1 and their sample images are shown in Figures 6 and 7.

Table 1. Dataset summary.

	North Campus	**Norland**
Environment	University of Michigan's North Campus	Train ride
Collection tools	Segway robot	Train
No. of frames (Reference/Query)	501/501	3600/3600
Distance between adjacent images	5 m	20 m
Viewpoint variation	Severe	None
Illumination variation	Severe	Severe
Seasonal variation	Severe	Severe
Tolerance (frames)	1	1

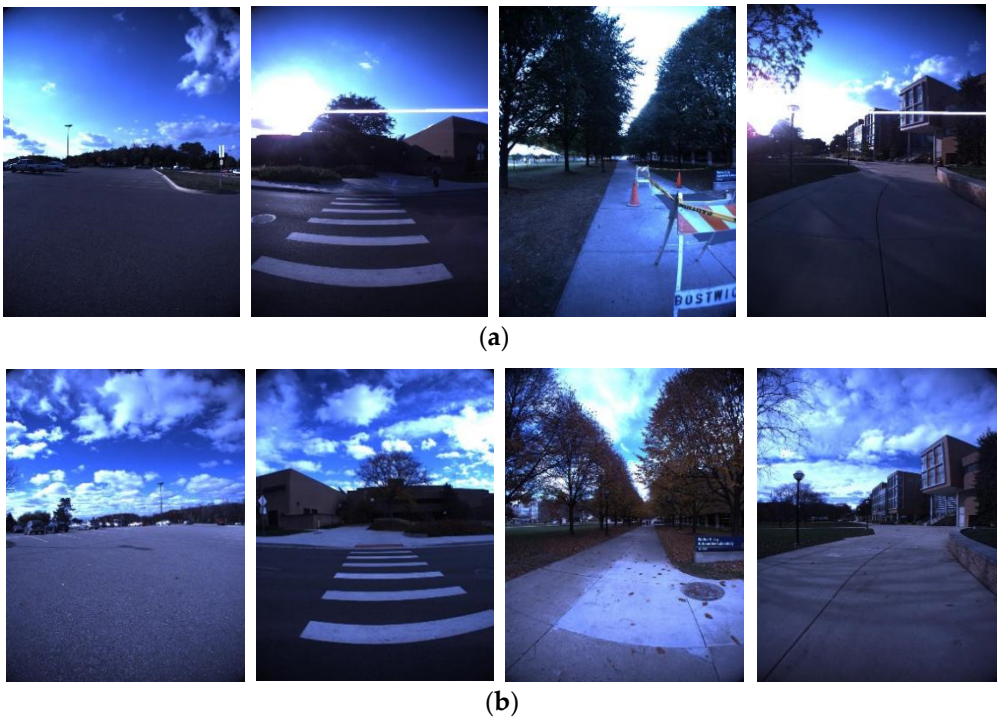


Figure 6. Image samples from the North Campus Dataset. (**a**) North-Summer-Left; (**b**) North-Autumn-Right.

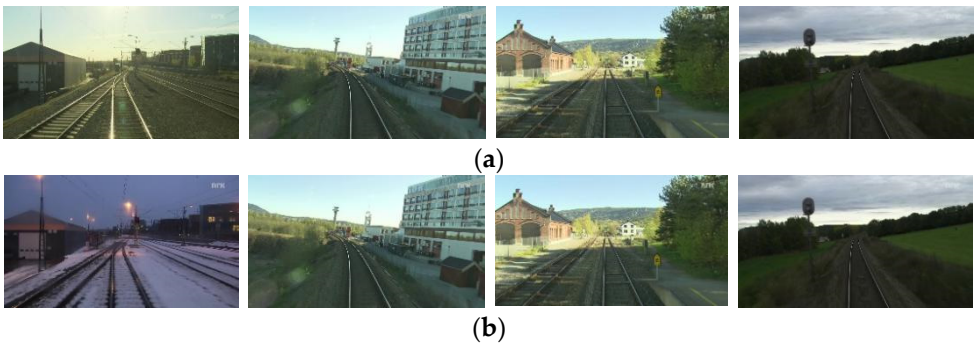


Figure 7. Image samples from the Norland Dataset. (**a**) Norland-Spring; (**b**) Norland-Winter.

3.1.1. North Campus Dataset

The North Campus Dataset is a large scale, long-term autonomy dataset for robotics research collected the University of Michigan's North Campus over 15 months. The dataset consists of 27 sequences which repeatedly explore the campus both indoor and outdoor on different trajectories across seasons, each containing dynamic obstacles, viewpoint variation, illumination variation, seasonal and weather changes, and long-term structural changes caused by construction. We used the summer sequence for reference and the autumn sequence for query. Figure 6 gives the image samples from the North Campus Dataset.

3.1.2. Norland Dataset

The Norland Dataset is the collection of four sequences of images from a 728 km trainway with seasonal environmental variation. Since the collection camera is fixed on the train head, there is no viewpoint variation. We used the spring sequence for reference and the summer for query. Image samples form Norland Dataset are shown in Figure 7.

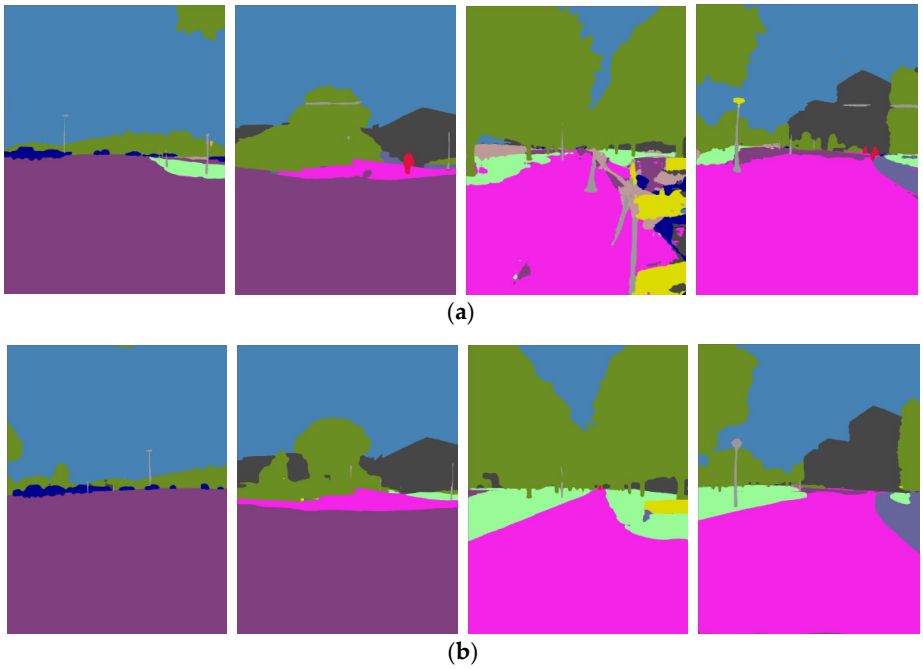
3.1.3. Performance Evaluations

We evaluated the recognition performance based on PR curve (precision-recall rate curve), matching time, and F1-score. The PR curve was used in the comparison experiment, and the matching time and F1-score were used in the ablation study. For each dataset, ground truth is the frame-level correspondence, and we set a tolerance of one frame. For each query, if the matched reference image was close enough to the correct reference image, it will be considered as a true positive match. For example, if the correct reference image is the kth image, then the $(k-1)$th, kth, and $(k+1)$ th reference image are all considered to be the true positive match to the query.

3.2. Experimental Setup

The semantic segmentation model is trained with the Cityscapes dataset [2] and then fine-tuned with the CMU-Seasons dataset [42]. The Cityscapes dataset includes 20 classes, then C is set as 20 and c is a value of 0–19.

The segmentations of sample images in Figures 6 and 7 are shown in Figures 8 and 9.



(**a**)

(**b**)

Figure 8. Segmentations of sample images from the North Campus Dataset: (**a**) North-Summer-Left; (**b**) North-Autumn-Right.

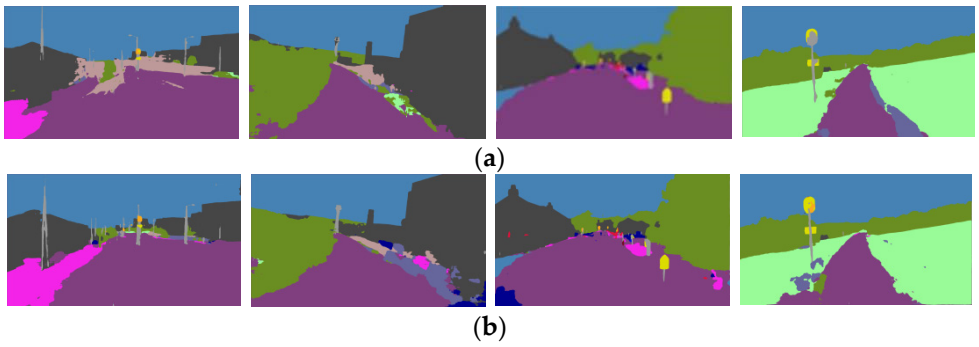


Figure 9. Segmentations of sample images from the Norland Dataset: **(a)** Norland-Spring; (**b**) Norland-Winter.

L is set as 3 representing three main static semantic classes of road, building, and vegetation in the images of query and reference datasets. Then, the hybrid image descriptor is simplified as $H = \left\langle \dot{H}_{road} + \dot{H}_{building} + \dot{H}_{vegetation} \right\rangle$, and $\dot{H}_{road}$, $\dot{H}_{building}$, $\dot{H}_{vegetation}$ refer to L2-normalized H_{road}, $H_{building}$, $H_{vegetation}$ respectively. Through the fowling ablation study, we take the *Candidates* n for 10.

We set $P = 64$ in the fine matching and kept the even coefficients of the wavelet transforms, which are redundant. Through this, we obtained a 128-dimension vector of the edge descriptor.

3.3. Ablation Study (Effects for Hierarchy and Candidates)

In order to study the effectiveness of hierarchy strategy and the number of *Candidates* in our method, we conducted 3.3.1 (the number of *Candidates*) and 3.3.2 (hierarchy or single), two ablation experiments on two datasets.

3.3.1. The Number of Candidates

To analyze the influence of the number of *Candidates(n)* on the whole method, matching time and F1-Score were adopted as the performance indicators. Note that matching time here refers to the time of coarse matching and fine matching but not the time of semantic segmentation of query and reference datasets.

We set the number of *Candidates* to 5, 10, 15, 20, 25, 30, and 50, respectively. The results of matching time and F1-Score with different number of *Candidates* are shown in Figures 10 and 11.

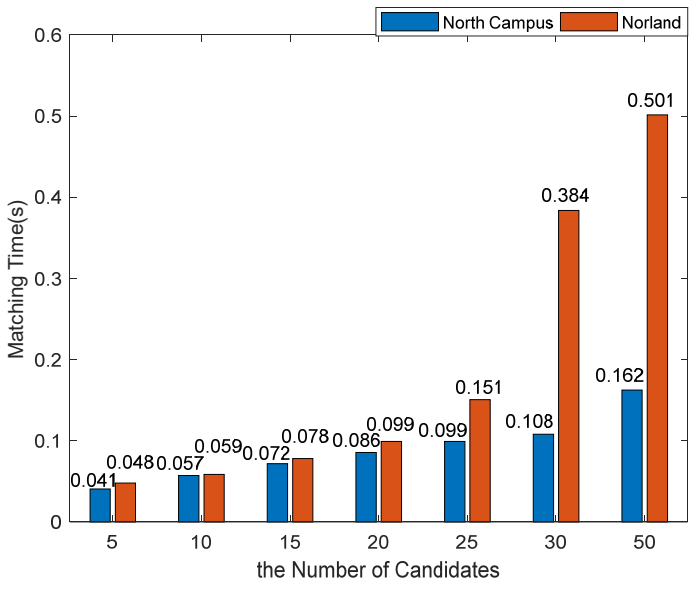


Figure 10. Matching time comparison of different number of *Candidates*.

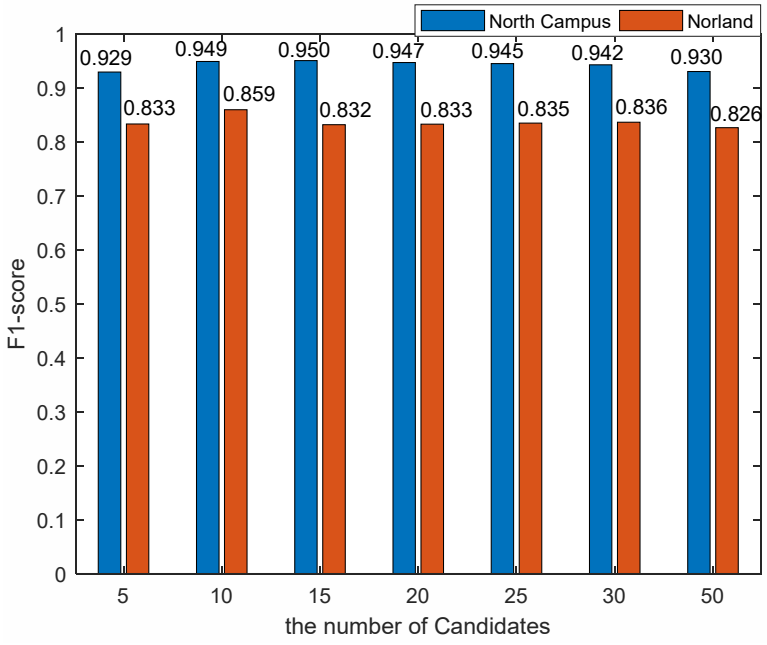


Figure 11. F1-score comparison of different number of *Candidates*.

The results in Figure 10 show that matching time of the North Campus dataset is lower than that of the Norland dataset on the whole. This is because the size of the two datasets is significantly different. The latter is 7 times more than that of the former, so the matching time of query on the Norland dataset is higher. Moreover, the matching time of the Norland Dataset increases greatly when the number of Candidates is 30 and 50. However, the maximum is only 0.501 s, which meets the real-time requirements. To sum up, matching time under 25 can be suitable for these two datasets.

As can be seen from Figure 11, for each dataset, there is little difference in the F1-Score of different *Candidates*. However, the F1-score of the North Campus Dataset is higher than that of the Norland Dataset on the whole. This indicates that our method is robust to severe viewpoint changes and image appearance changes.

Taking account of matching time and F1-score, we find that the effectiveness is better when the number of Candidates n is 10 or 15. Finally, in the comparison experiment, we took n for 10.

3.3.2. Hierarchy or Single

To compare the performance of hierarchical place recognition method, coarse matching only, and fine matching only, we conducted the Hierarchy or Single experiments. F1-Score was adopted as the performance indicator.

Note that coarse matching only means that we get the final best match for the query just through a coarse matching. Fine matching only means that we get the final best match for the query just by fine matching.

According to the result of ablation 3.3.1, we compared our hierarchical method with the *Candidates* n to be 10 and 15, respectively, coarse matching only and fine matching only on the North Campus Dataset and the Norland Dataset. The results are shown in Table 2.

The results show that the performance of hierarchical strategy with *Candidates* 10 and 15 is better than that of coarse matching only, and fine matching only on both two datasets. It reveals that our hierarchical place recognition is efficient. Comparing with using a single strategy, the hierarchical strategy behaves better.

Table 2. F1-scores of different strategies.

	The North Campus Dataset	The Norland Dataset
Hierarchical strategy with Candidates 10.	0.94856	0.85944
Hierarchical strategy with Candidates 15.	0.95046	0.83152
Coarse matching only	0.85045	0.75051
Fine matching only	0.90979	0.80217

3.4. Comparison with the State-of-Art Methods

We conducted experiments to evaluate the performance of place recognition by comparing PR curves of the following single-image-based baseline methods:

FabMap [4]: A classical method for appearance-based VPR based on Bag-of-Words model;

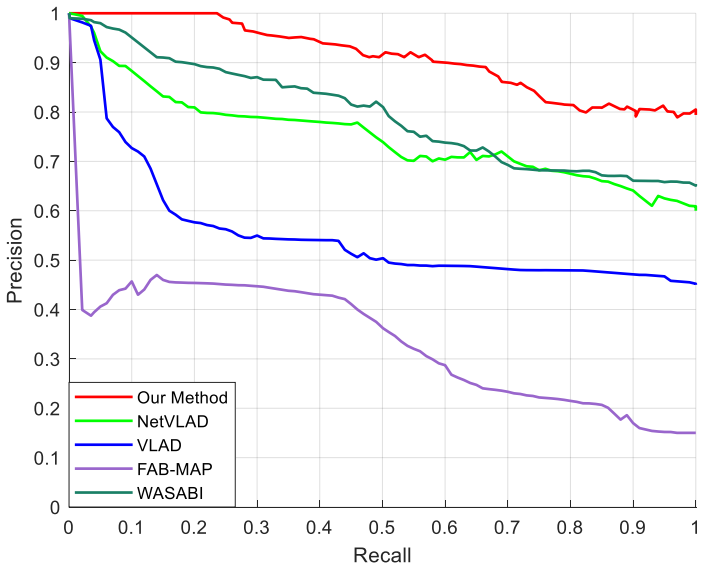
VLAD [13]: A large-scale image-based place recognition model. It can be used for place recognition and realize good performance on many datasets;

NetVLAD [20]: A viewpoint-robust CNN model for VPR, which can achieve great performance on most datasets;

WASABI [31]: A novel image-based place recognition model across seasons from semantic edge description on bucolic environments such as scenes with low texture and little semantic content.

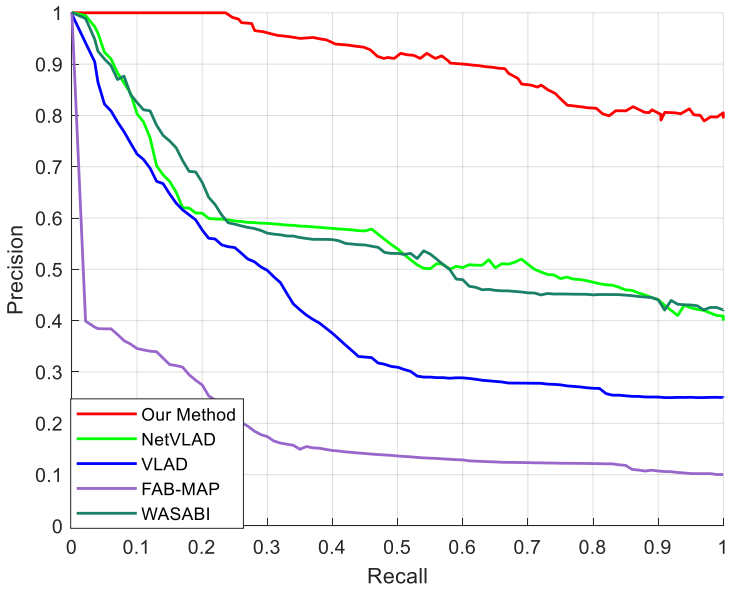
The results are shown in Figure 12. Figure 12a shows the results of experiments conducted on the North Campus Dataset, which involves severe viewpoint variations and environmental condition variations. Figure 12b shows the results of experiments conducted on the Norland Dataset, which involves severe appearance variations. The method that we proposed (red line) obtains the best performance. We think this is because our method utilizes both the mid-level convolutional features and the higher-level semantic features in the coarse matching, thus our method is robust to the viewpoint changes and environmental conditions. Moreover, the fine matching further improves the accuracy.

The results indicate that the method that we proposed is robust to viewpoint-variant and appearance-variant conditions.



(**a**)

Figure 12. *Cont.*



(**b**)

Figure 12. Precision-recall curves of our method and baseline methods: (**a**) North Summer-Left (reference) and North Autumn-Right (query); (**b**) Norland-Spring (reference) and Norland-Winter (query).

3.5. Runtime Analysis

We implemented the proposed system in two steps: (1) semantic segmentation and (2) coarse matching and fine matching. We called the step (2) the matching process, and experiments were done on an NVIDIA 1090Ti GPU. The results are shown in Figure 10. For a single image, it takes approximately 0.059 s to achieve matching with the *Candidates* 10. Even when the *Candidates* number is 50, the time of matching is 0.501 s for a reference image sequence with 3600 images. We believe that our method has the potential to satisfy real-time demands.

4. Discussion

We presented a coarse-to-fine visual place recognition pipeline, and done experiments on two benchmark databases with many images from a wide variety of seasonal environments to study whether our method adapts to variations in viewpoint and appearance. We compared our method with state-of-art place recognition algorithms and demonstrated its superior performance. Our proposed method can be used in loop-closure and localization, and it performs well especially for the scenes with seasonal environmental changes and long-term conditional changes. However, it is important to note that our method relies on semantic segmentation. Thus, the effectiveness of semantic segmentation has a great influence on our method and the performance of the computer also affects the efficiency of our method.

5. Conclusions

In this article, we proposed a coarse-to-fine hierarchical place recognition based on semantic-aggregation. Specifically, we aggregate the mid-level convolutional feature and high-level semantic feature in the coarse matching, while associating semantic edges in fine matching. The experimental results show that our method significantly improves the performance, exhibiting strong robustness against variations in viewpoint and appearance simultaneously. It outperforms the state-of-art single-image-based methods on two representative datasets while showing good computational efficiency. In the future, we will study how to improve our method, making it adapted to sequence-image-based place recognition.

Author Contributions: Conceptualization, B.C. and X.S.; validation, X.S.; formal analysis, B.C. and X.S.; investigation, X.S.; data curation, X.S.; writing—original draft preparation, X.S.; writing—review and editing, B.C., X.S. and H.S.; supervision, T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Foundation of Hubei Key Laboratory of Intelligent of Robot (Wuhan Institute of Technology), grant number HBIR202009, and the Key Laboratory of Hunan Province for New Retail Virtual Reality Technology (2017TP1026).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sattler, T.; Leibe, B.; Kobbelt, L. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *IEEE Trans. Patt. Anal. Mac. Intell.* **2016**, *39*, 1744–1756.
2. Sattler, T.; Maddern, W.; Toft, C. Benchmarking 6dof outdoor visual localization in changing conditions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8601–8610.
3. Brahmbhatt, S.; Gu, J.; Kim, K. Geometry-aware learning of maps for camera localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2616–2625.
4. Cummins, M.; Newman, P. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *Int. J. Robot. Res.* **2008**, *27*, 647–665. [CrossRef]
5. Galvez-Lpez, D.; Tardos, J.D. Bags of Binary Words for Fast Place Recognition in Image Sequences. *IEEE Trans. Robot.* **2012**, *28*, 1188–1197. [CrossRef]
6. Lowry, S.; Sunderhauf, N.; Newman, P.; Leonard, J.J.; Cox, D.; Corke, P.; Milford, M.J. Visual Place Recognition: A Survey. *IEEE Trans. Robot.* **2016**, *32*, 1–19. [CrossRef]
7. Milford, M.J.; Wyeth, G.F. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation (ICRA), Melbourne, Australia, 9–13 July 2012; pp. 1643–1649.
8. Sattler, T.; Havlena, M.; Schindler, K.; Pollefeys, M. Large-scale location recognition and the geometric burstiness problem. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1582–1590.
9. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
10. Bay, H.; Ess, A.; Tuytelaars, T. Speeded-up robust features (SURF). *Comput. Vis. Image Understand.* **2008**, *110*, 346–359. [CrossRef]
11. Nicosevici, T.; Garcia, R. Automatic visual bag-of-words for online robot navigation and mapping. *IEEE Trans. Robot.* **2012**, *28*, 886–898. [CrossRef]
12. Milford, M.; Scheirer, W.; Vig, E. Condition-invariant top-down visual place recognition. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–5 June 2014; pp. 5571–5577.
13. Amato, G.; Bolettieri, P.; Falchi, F. Large scale image retrieval using vector of locally aggregated descriptors. In Proceedings of the International Conference on Similarity Search and Applications, A Coruña, Spain, 2–4 October 2013; Springer; Berlin/Heidelberg, Germany, 2013; pp. 245–256.
14. Cadena, C.D.; Galvez-Lopez, D.; Tardos, J.D. Robust Place Recognition With Stereo Sequences. *IEEE Trans. Robot.* **2012**, *28*, 871–885. [CrossRef]
15. Lu, F.; Chen, B.; Guo, Z. Visual sequence place recognition with improved dynamic time warping. In Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 4–6 November 2019; pp. 1034–1041.
16. Chen, B.; Yuan, D.; Liu, C.; Wu, Q. Loop closure detection based on multi-scale deep feature fusion. *Appl. Sci.* **2019**, *9*, 1120. [CrossRef]
17. Wang, T.H.; Huang, H.J.; Lin, J.T.; Hu, C.W.; Zeng, K.H.; Sun, M. Omnidirectional cnn for visual place recognition and navigation. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 2341–2348.
18. Chen, Z.; Lin, O.; Jacobson, A. Convolutional neural network-based place recognition. *arXiv* **2014**, arXiv:1411.1509.
19. Sunderhauf, N.; Shirazi, S.; Dayoub, F. On the performance of convnet for place recognition. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–3 October 2015; pp. 4297–4304.
20. Arandjelovic, R.; Gronat, P.; Torii, A. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5297–5307.
21. Chen, Z.; Maffra, F.; Sa, I. Only look once, mining distinctive landmarks from convnet for visual place recognition. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 9–16.

22. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
23. Ronneberger, O.; Fischer, P.; Brox, T. U-net:Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer; Cham, Switzerland, 2015; pp. 234–241.
24. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet:A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Patt. Anal. Mach. Int.* **2017**, *39*, 2481–2495. [CrossRef]
25. Chen, L.C.; Papandreou, G.; Kokkinos, I. Deeplab:Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Patt. Anal. Mach. Int.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
26. Chen, L.C.; Papandreou, G.; Schroff, F. Rous Convolution of Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
27. Zhao, H.; Shi, J.; Qi, X. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
28. He, K.; Gkioxari, G.; Yuan, P. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
29. Garg, S.; Suenderhauf, N.; Milford, M. Don't look back:Robustifying place categorization for viewpoint-and condition-invariant place recognition. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 3645–3652.
30. Garg, S.; Suenderhauf, N.; Milford, M. Semantic-geometric visual place recognition: A new perspective for reconciling opposing views. *Int. J. Robot. Res.* **2019**. [CrossRef]
31. Benbihi, A.; Arravechia, S.; Geist, M. 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), Paris, France, 31 May–31 August 2020; pp. 3032–3038.
32. Maohai, L.; Lining, S.; Qingcheng, H. Robust omnidirectional vision based mobile robot hierarchical localization and autonomous navigation. *Inf. Technol. J.* **2011**, *10*, 29–39. [CrossRef]
33. Garcia-Fidalgo, E.; Ortiz, A. Hierarchical place recognition for topological mapping. *IEEE Trans. Robot.* **2017**, *33*, 1061–1074. [CrossRef]
34. Hausler, S.; Milford, M. Hierarchical multi-process fusion for visual place recognition. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 30 May–5 June 2020; pp. 3327–3333.
35. Larsson, M.; Stenborg, E.; Hammarstrand, L. A cross-season correspondence dataset for robust semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9532–9542.
36. He, K.; Zhang, X.; Ren, S. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
37. Chen, L.C.; Papandreou, G.; Kokkinos, I. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
38. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
39. Chuang, C.H.; Kuo, C. Wavelet descriptor of planar curves: Theory and applications. *IEEE Trans. Image Proc.* **1996**, *5*, 56–70. [CrossRef]
40. Carlevaris-Bianco, N.; Ushani, A.K.; Eustice, R.M. University of Michigan North Campus long-term vision and lidar dataset. *Int. J. Robot. Res.* **2016**, *35*, 1023–1035. [CrossRef]
41. Olid, D.; Fácil, J.M.; Civera, J. Single-view place recognition under seasonal changes. *arXiv* **2018**, arXiv:1808.06516.
42. Cordts, M.; Omran, M.; Ramos, S. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.