applied sciences

MDPI

Article

Variational Bayesian Approach to Condition-Invariant Feature Extraction for Visual Place Recognition

Junghyun Oh [1] and Gyuho Eoh [2,*]

[1] Department of Robotics, Kwangwoon University, Seoul 01897, Korea; jhyunoh@kw.ac.kr
[2] Industrial AI Research Center, Chungbuk National University, Cheongju 28116, Korea
* Correspondence: gyuho.eoh@cbnu.ac.kr

Abstract: As mobile robots perform long-term operations in large-scale environments, coping with perceptual changes becomes an important issue recently. This paper introduces a stochastic variational inference and learning architecture that can extract condition-invariant features for visual place recognition in a changing environment. Under the assumption that a latent representation of the variational autoencoder can be divided into condition-invariant and condition-sensitive features, a new structure of the variation autoencoder is proposed and a variational lower bound is derived to train the model. After training the model, condition-invariant features are extracted from test images to calculate the similarity matrix, and the places can be recognized even in severe environmental changes. Experiments were conducted to verify the proposed method, and the experimental results showed that our assumption was reasonable and effective in recognizing places in changing environments.

Keywords: place recognition; localization; deep learning; mobile robots; auto-encoder; SLAM

Citation: Oh, J.; Eoh, G. Variational Bayesian Approach to Condition-Invariant Feature Extraction for Visual Place Recognition. *Appl. Sci.* **2021**, *11*, 8976. https://doi.org/10.3390/app11198976

Academic Editor: Antonio Fernández-Caballero

Received: 30 August 2021
Accepted: 23 September 2021
Published: 26 September 2021

1. Introduction

Autonomous robots operating over long periods of time, such as days, weeks, or months, face a variety of environmental changes over time. As the environment changes, robots should recognize places using their visual sensors, which is called long-term visual place recognition. It is an essential component for achieving long-term simultaneous localization and mapping (SLAM) and autonomous navigation [1]. One of the major problems in long-term visual place recognition is the appearance change problem caused by factors such as time of day or weather conditions [2].

To solve the appearance change problem in visual place recognition, global descriptors that can describe the whole-image have widely used [3,4]. Compared to local features such as SIFT [5] and SURF [6], global descriptors are not only robust to illumination changes, but also require less computation since they do not require a keypoint detection phase [1]. The classic hand-crafted global descriptors such as HOG [7] or gist [8,9] showed higher place recognition performance than the existing local descriptors in a changing environment [3,4]. However, hand-crafted descriptors have inherent limitations in generalization performance since features are extracted according to predefined parameters.

Recently, features from deep learning structures have proven to have superior generalization performances than existing hand-crafted methods. In particular, a deep convolutional neural network (CNN), a kind of neural network, is a structure that have shown excellent performance in image recognition and classification [10]. A variety of structures using CNNs have been widely used in visual place recognition [11–14]. A sequence of image features using CNNs was used to find the same places between different seasons in [15]. Sünderhauf et al. evaluated CNNs features from each layer of pretrained AlexNet [10] for visual place recognition in a changing environment. Another deep learning structure, the autoencoder, has been also used for visual place recognition because the output of each layer can be used as an image descriptor. Oh and Lee [16] used a deep convolutional autoen-

coder (CAE) for feature extraction, and Park [17] proposed an illumination-compensated CAE for robust place recognition.

In this paper, we propose a novel feature extraction method based on variational autoencoders (VAEs) [18]. It is one of the popular models for unsupervised representation learning, and showed outstanding performance in feature learning [19,20]. It consists of a standard autoencoder component, and can approximate Bayesian inference for latent variable models. To obtain robust performances in a changing environment, we assume that the image $\mathbf{x}$ is generated from the latent variable $\mathbf{z}$, and this latent representation is divided into the condition-invariant feature $\mathbf{z}_p$ and the condition-sensitive feature $\mathbf{z}_c$. To find the same places from different conditions, comparing the condition-invariant features improves the performance of place recognition. The proposed procedure is shown in Figure 1.

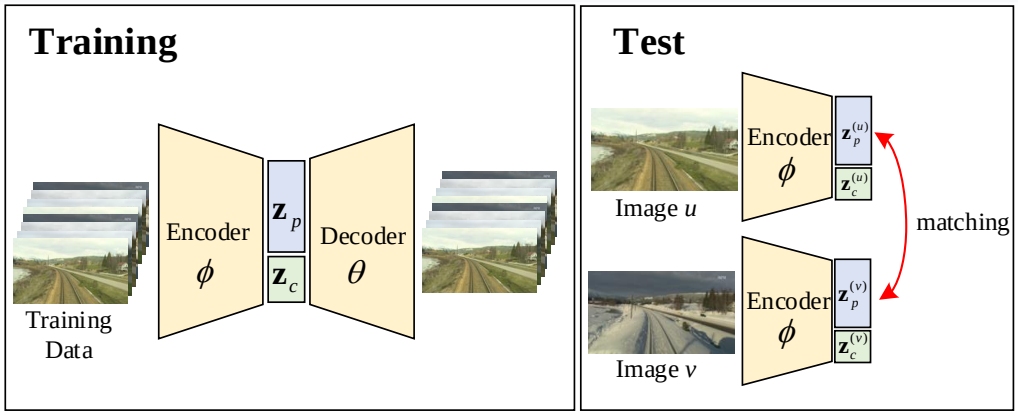

Figure 1. The proposed model of VAE for condition-invariant feature extraction in a changing environment. After training the model, the same place can be recognized by extracting encoded features from images obtained in different environments.

Our paper is organized as follows. Section 2 explains the basic preliminaries of VAEs. The proposed structures for feature extraction using the context information is explained in Section 3. Then, the robot localization using the extracted condition-invariant feature is discussed in Section 4. Section 5 presents the validation of the proposed method through publicly available datasets with other algorithms. Finally, Section 6 concludes the paper.

2. Preliminaries

Let us consider the dataset $\mathbf{X}$ consisting of N images $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}\}$. The assumption of the generative model is that the observed images are generated by some stochastic process, involving an unobserved random variable $\mathbf{z}$. To be specific, the latent representation $\mathbf{z}^{(i)}$ is generated from a prior distribution $p(\mathbf{z})$, and the image $\mathbf{x}^{(i)}$ is generated from a conditional distribution $p_\theta(\mathbf{x}|\mathbf{z})$ where θ is the generative model parameter.

To efficiently approximate posterior inference of the latent variable $\mathbf{z}$ given an observed value $\mathbf{x}$, a recognition model $q_\phi(\mathbf{z}|\mathbf{x})$ is introduced where ϕ is the recognition model parameter. This model is an approximation to the intractable true posterior $p_\theta(\mathbf{x}|\mathbf{z})$, and also referred as a probabilistic *encoder*. Instead of encoding an input image $\mathbf{x}$ as a single vector, the encoder produces a probabilistic distribution of the compressed feature $\mathbf{z}$ over the latent space. Similarly, $p_\theta(\mathbf{x}|\mathbf{z})$ is a probabilistic *decoder* since given a latent feature $\mathbf{z}$ it produces a probabilistic distribution over the possible corresponding values of $\mathbf{x}$.

The VAE is a structure that implements an encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and a decoder $p_\theta(\mathbf{x}|\mathbf{z})$ as a neural network as shown in Figure 2.

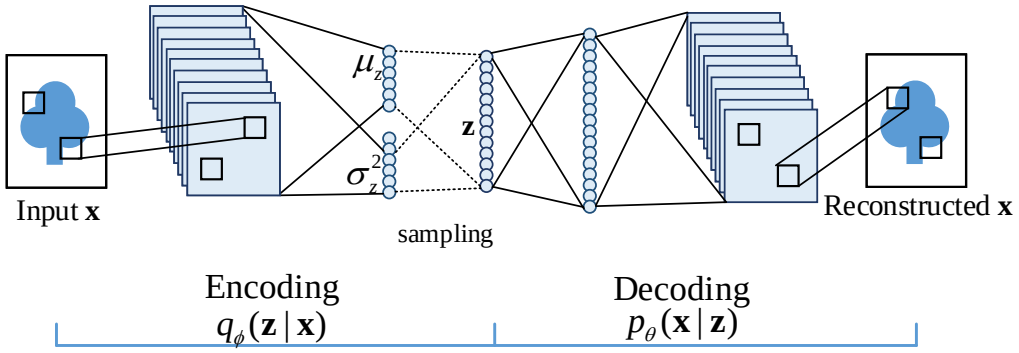

Figure 2. The structure of the vanilla VAE composed of the encoder and the decoder.

Then, parameters ϕ and θ become the weights of the neural network. The objective is to find the ϕ and θ maximizing the variational lower bound $\mathcal{L}(\theta, \phi; \mathbf{x})$ on the marginal likelihood [18] as the following:

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) \quad (1)$$

where $D_{KL}(\cdot)$ stands for the Kullback–Leibler divergence, which measures the difference between two probability distributions. The objective function consists of a reconstruction likelihood and a regularization term. The prior distribution $p_\theta(\mathbf{z})$ is usually set to a Gaussian distribution so that the reparameterization trick can be used to train the network [18].

After training the VAE, it can compress the input image to the low-dimensional latent vector $\mathbf{z}$. Since the encoded vector $\mathbf{z}$ contains the information of the whole input image, it can be used as a global descriptor for comparing similarities between images [19].

3. Proposed VAE Using Context Information

Although the compressed vector $\mathbf{z}$ can be used as a useful global descriptor, it is insufficient to cope with environmental changes. To find the same place obtained from different environments, external factors such as weather or seasonal changes should be removed from the vector $\mathbf{z}$. If the vector $\mathbf{z}$ is divided into the condition-invariant feature $\mathbf{z}_p$ and the condition-sensitive feature $\mathbf{z}_c$, we would be able to reliably distinguish places even in changing environments using only the condition-invariant feature $\mathbf{z}_p$.

To achieve this goal, we assume that observed images are affected by both structural information $\mathbf{p}$ such as unique landmarks, and context information $\mathbf{c}$ due to environmental changes such as light or weather changes. Since structural information is robust and context information is sensitive to environmental changes, each information is contained in the condition-invariant feature $\mathbf{z}_p$ and the condition-sensitive feature $\mathbf{z}_c$, respectively. To divide the latent feature $\mathbf{z}$ into $\mathbf{z}_p$ and $\mathbf{z}_c$, we propose a structure for generating the context vector c from $\mathbf{z}_c$ and the image from both $\mathbf{z}_p$ and $\mathbf{z}_c$. Therefore, the generative model is changed from $p_\theta(\mathbf{x}|\mathbf{z})$ to $p_{\theta,\varphi}(\mathbf{x}, \mathbf{c}|\mathbf{z}_p, \mathbf{z}_c)$, and is factorized as the following:

$$p_{\theta,\varphi}(\mathbf{x}, \mathbf{c}|\mathbf{z}_p, \mathbf{z}_c) = p_\theta(\mathbf{x}|\mathbf{z}_p, \mathbf{z}_c) \cdot p_\varphi(\mathbf{c}|\mathbf{z}_c) \quad (2)$$

where θ and φ are parameters of the generative model to generate $\mathbf{x}$ and $\mathbf{c}$, respectively. The comparison between the existing and proposed generative model is shown in Figure 3.

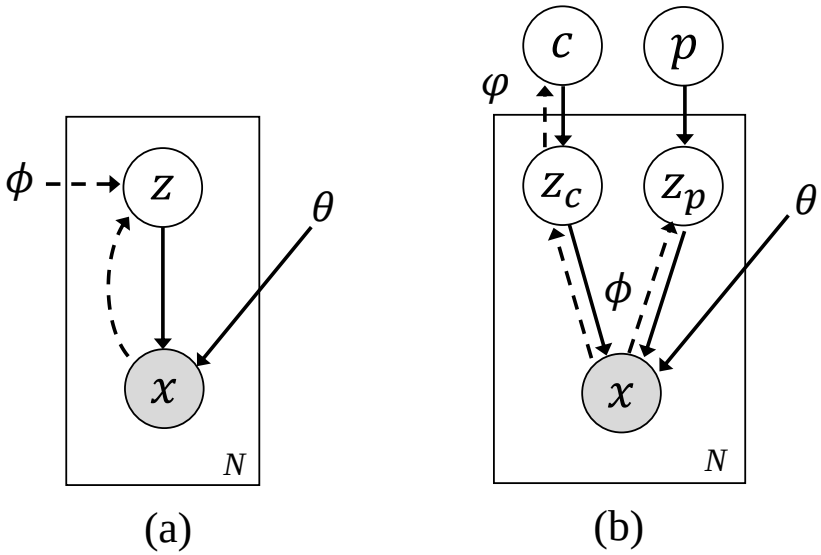

Figure 3. The comparison between (**a**) existing model and (**b**) proposed graphical models for data generation. Solid lines denote the generative model and dashed lines denote the recognition model. The proposed model assumes that images are generated from the condition-invariant feature $\mathbf{z}_p$ and the condition-sensitive feature $\mathbf{z}_c$.

Then, the variational lower bound is also modified from $\mathcal{L}(\theta, \phi; \mathbf{x})$ to $\mathcal{L}(\theta, \phi, \varphi; \mathbf{x}, \mathbf{c})$ on the marginal likelihood as follows:

$$\begin{aligned}\mathcal{L}(\theta, \phi, \varphi; \mathbf{x}, \mathbf{c}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_{\theta,\varphi}(\mathbf{x}, \mathbf{c}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_{\theta,\varphi}(\mathbf{z})) \\ &= \mathbb{E}_{q_\phi(\mathbf{z}_p, \mathbf{z}_c|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_p, \mathbf{z}_c) + \log p_\varphi(\mathbf{c}|\mathbf{z}_c)] \\ &\quad - D_{KL}(q_\phi(\mathbf{z}_p, \mathbf{z}_c|\mathbf{x})||p_\theta(\mathbf{z}_p, \mathbf{z}_c)p_\varphi(\mathbf{z}_c))\end{aligned} \tag{3}$$

In order to learn the probability distributions, our proposed structure named *C-VAE* is shown in Figure 4. The encoding part is considered as the inference model $q_\phi(\mathbf{z}_p, \mathbf{z}_c|\mathbf{x})$, and the decoding part is the generative model $p_\theta(\mathbf{x}|\mathbf{z}_p, \mathbf{z}_c)$ and $p_\varphi(\mathbf{c}|\mathbf{z}_c)$.

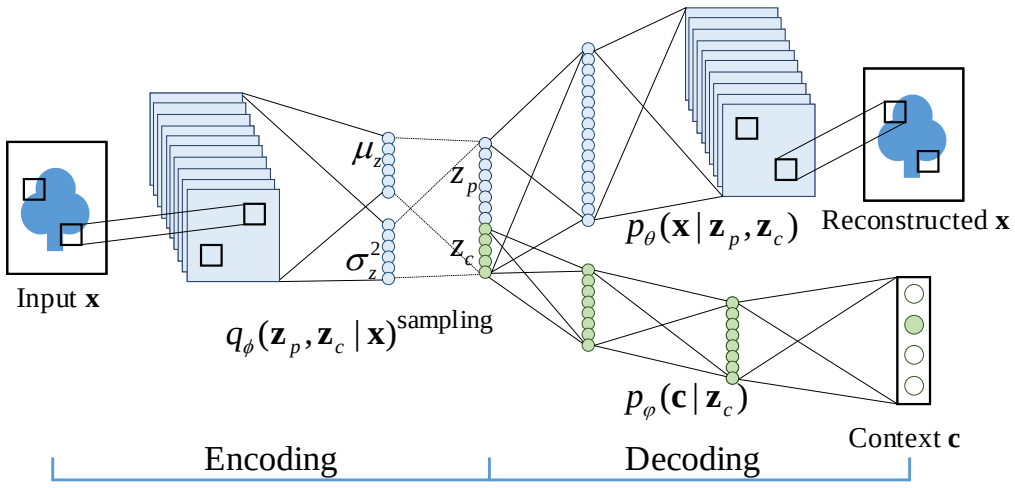

Figure 4. The structure of the C-VAE for feature extraction in changing environments.

A detailed examination of this structure reveals the following characteristics in comparison with the existing VAE. The reconstruction of the input image **x** is the same as the existing structure. The difference is that $\mathbf{z}_c$, a subset of $\mathbf{z}$, is used not only to reconstruct **x**, but also to create the context vector **c**. During the learning process, information that

is sensitive to environmental influences is concentrated in $\mathbf{z}_c$, and condition-invariant information is compressed into $\mathbf{z}_p$. Therefore, $\mathbf{z}_p$ can be used as a feature of an image which is robust to environmental changes.

If not only context information $\mathbf{c}$ but also structural information $\mathbf{p}$ is provided, we propose a model named *CP-VAE* as shown in Figure 5, which improves the independence between $\mathbf{z}_p$ and $\mathbf{z}_c$ of the previous model. The generative model is modified to $p_{\theta,\varphi,\psi}(\mathbf{x},\mathbf{p},\mathbf{c}|\mathbf{z}_p,\mathbf{z}_c)$, and factorized as the following:

$$p_{\theta,\psi,\varphi}(\mathbf{x},\mathbf{c}|\mathbf{z}_p,\mathbf{z}_p,\mathbf{z}_c) = p_\theta(\mathbf{x}|\mathbf{z}_p,\mathbf{z}_c)\cdot p_\psi(\mathbf{p}|\mathbf{z}_p)\cdot p_\varphi(\mathbf{c}|\mathbf{z}_c) \tag{4}$$

where θ, ψ, and φ are parameters of the generative model to generate $\mathbf{x}$, $\mathbf{p}$ and $\mathbf{c}$, respectively. The variational lower bound is also modified as follows:

$$\begin{aligned}\mathcal{L}(\theta,\phi,\psi,\varphi;\mathbf{x},\mathbf{p},\mathbf{c}) \\ = \mathbb{E}_{q_\phi(\mathbf{z}_p,\mathbf{z}_c|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_p,\mathbf{z}_c) + \log p_\psi(\mathbf{p}|\mathbf{z}_p) + \log p_\varphi(\mathbf{c}|\mathbf{z}_c)] \\ - D_{KL}(q_\phi(\mathbf{z}_p,\mathbf{z}_c|\mathbf{x})||p_\theta(\mathbf{z}_p,\mathbf{z}_c)p_\psi(\mathbf{z}_p)p_\varphi(\mathbf{z}_c))\end{aligned} \tag{5}$$

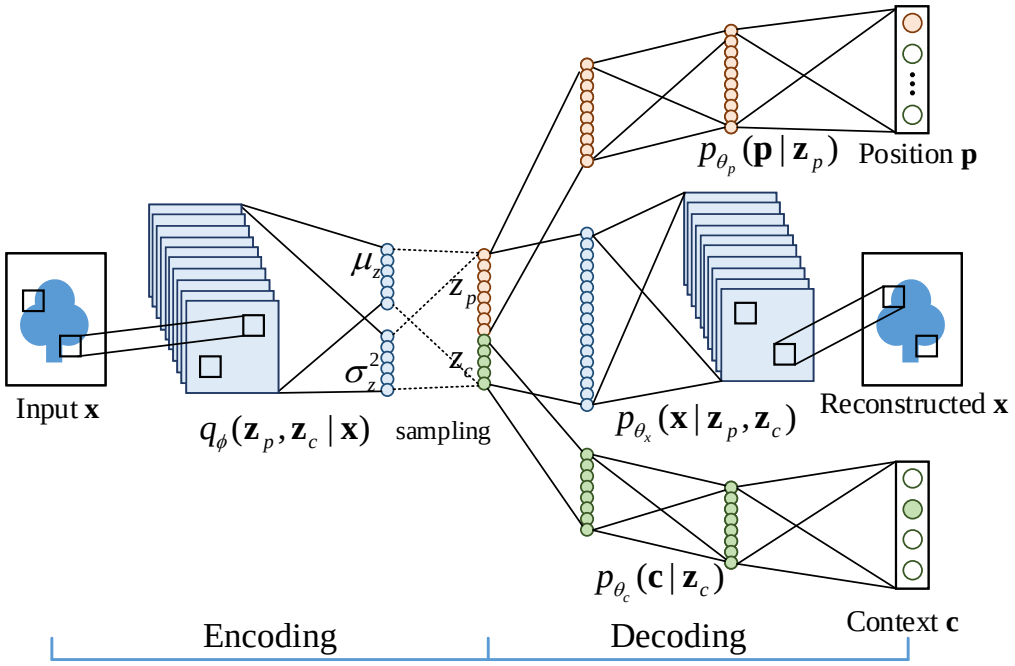

Figure 5. The structure of the CP-VAE for feature extraction in changing environments.

The difference from the previous model is that $\mathbf{z}_p$ generates not only the image $\mathbf{x}$, but also the position information vector $\mathbf{p}$. Since $\mathbf{z}_p$ generates the structural information vector $\mathbf{p}$, the independence between $\mathbf{z}_p$ and $\mathbf{z}_c$ is enhanced, and the more robust condition-invariant feature $\mathbf{z}_p$ can be extracted to recognize places under substantial environmental changes. However, this model has a limitation in that it requires a fairly strong assumption that the training data are aligned with the same places in order to obtain the position vector information $\mathbf{p}$.

4. Robot Localization Using Condition-Invariant Features

After training the model, the encoding part of the proposed structure can be used to extract the condition-invariant feature $\mathbf{z}_p$ from the image. If there are two image sequences ${}^u\mathbf{X} = \{{}^u\mathbf{x}^{(1)}, {}^u\mathbf{x}^{(2)}, ..., {}^u\mathbf{x}^{(M)}\}$ and ${}^v\mathbf{X} = \{{}^v\mathbf{x}^{(1)}, {}^v\mathbf{x}^{(2)}, ..., {}^v\mathbf{x}^{(N)}\}$ from different environments u and v, we can extract feature sequence ${}^u\mathbf{Z} = \{{}^u\mathbf{z}_p^{(1)}, {}^u\mathbf{z}_p^{(2)}, ..., {}^u\mathbf{z}_p^{(M)}\}$ and ${}^v\mathbf{Z} = \{{}^v\mathbf{z}_p^{(1)}, {}^v\mathbf{z}_p^{(2)}, ..., {}^v\mathbf{z}_p^{(N)}\}$, respectively. Then, the similarity matrix $S \in \mathbb{R}^{M\times N}$ can be

constructed from the affinity score between features. The component of the S is the affinity score s_{ij} between ${}^{u}\mathbf{z}_p^{(i)}$ and ${}^{v}\mathbf{z}_p^{(j)}$, where $1 \leq i \leq M$ and $1 \leq j \leq N$. It is calculated using the cosine similarity as follows:

$$s_{ij} = \frac{{}^{u}\mathbf{z}_p^{(i)} \cdot {}^{v}\mathbf{z}_p^{(j)}}{\|{}^{u}\mathbf{z}_p^{(i)}\| \|{}^{v}\mathbf{z}_p^{(j)}\|} \tag{6}$$

The affinity score s_{ij} has a value between $[0, 1]$, and the closer it is to 1, the higher the probability of the same place. From the similarity matrix S, we can find the correspondence between the query sequence ${}^{v}\mathbf{X}$ and the database sequence ${}^{u}\mathbf{X}$, and the location of the mobile robot can be successfully recognized.

5. Experimental Results

In this section, various experiments were performed to verify the performance of the proposed algorithm. We used the *Nordland dataset* [21,22], which comprises images of all seasons from four journeys on a 728 km train route across Norway, and the *KAIST dataset* [23], which includes six sequences in various illumination conditions: day, night, sunset, and sunrise. They are challenging datasets widely used for long-term place recognition because images between sequences show drastic appearance changes. In each sequence, 1600 images were used for training, and 6400 images were used as a test. All the images were resized to 224×224 pixels.

The output shape of the encoding part in our model is shown in Table 1. To effectively compress the data, several layers of convolutional and fully connected layers were used. Then, the output from the `sampling` layer is the latent feature z, and this feature is divided into $\mathbf{z}_p$ with 96 nodes and $\mathbf{z}_c$ with 32 nodes. The decoding part includes a part that reconstructs the input image $\mathbf{x}$ from the $\mathbf{z}_p$ and $\mathbf{z}_c$ similar to a typical VAE, and a part that generates a context vector $\mathbf{c}$ from the $\mathbf{z}_c$. Since the dataset has four seasons, the context vector $\mathbf{c}$ is defined as a four-dimensional one-hot encoding vector.

Table 1. The input and output shapes of the encoding part in our VAE model.

Layer	Input Size	Output Size
`conv1`	$224 \times 224 \times 3$	$112 \times 112 \times 32$
`conv2`	$112 \times 112 \times 32$	$56 \times 56 \times 64$
`conv3`	$56 \times 56 \times 64$	$28 \times 28 \times 64$
`conv4`	$28 \times 28 \times 64$	$14 \times 14 \times 128$
`conv5`	$14 \times 14 \times 128$	$7 \times 7 \times 128$
`fc6`	6272	4096
`fc7`	4096	2048
`fc8`	2048	1024
`fc9`	1024	512
`z_mean`	512	128
`z_var`	512	128
`sampling`	128, 128	128

The first experiment is a visualization test to confirm if the model has been trained to make $\mathbf{z}_p$ and $\mathbf{z}_c$ independent as intended. Let ${}^{u}\mathbf{x}$ and ${}^{v}\mathbf{x}$ be images obtained from different environments u and v, respectively. Then, we can extract the latent features ${}^{u}\mathbf{z} = \{{}^{u}\mathbf{z}_p, {}^{u}\mathbf{z}_c\}$ and ${}^{v}\mathbf{z} = \{{}^{v}\mathbf{z}_p, {}^{v}\mathbf{z}_c\}$ from each image respectively using the encoder of the trained model. Since the reconstructed image from the decoder is mainly affected by the condition-sensitive feature $\mathbf{z}_c$, not the condition-invariant feature $\mathbf{z}_p$, we can expect the reconstructed image from the combined feature $\{{}^{u}\mathbf{z}_p, {}^{v}\mathbf{z}_c\}$ will be $\mathbf{x}_v$. The results of combining $\mathbf{z}_p$ and $\mathbf{z}_c$ extracted from each sequence image are shown in Figures 6 and 7.

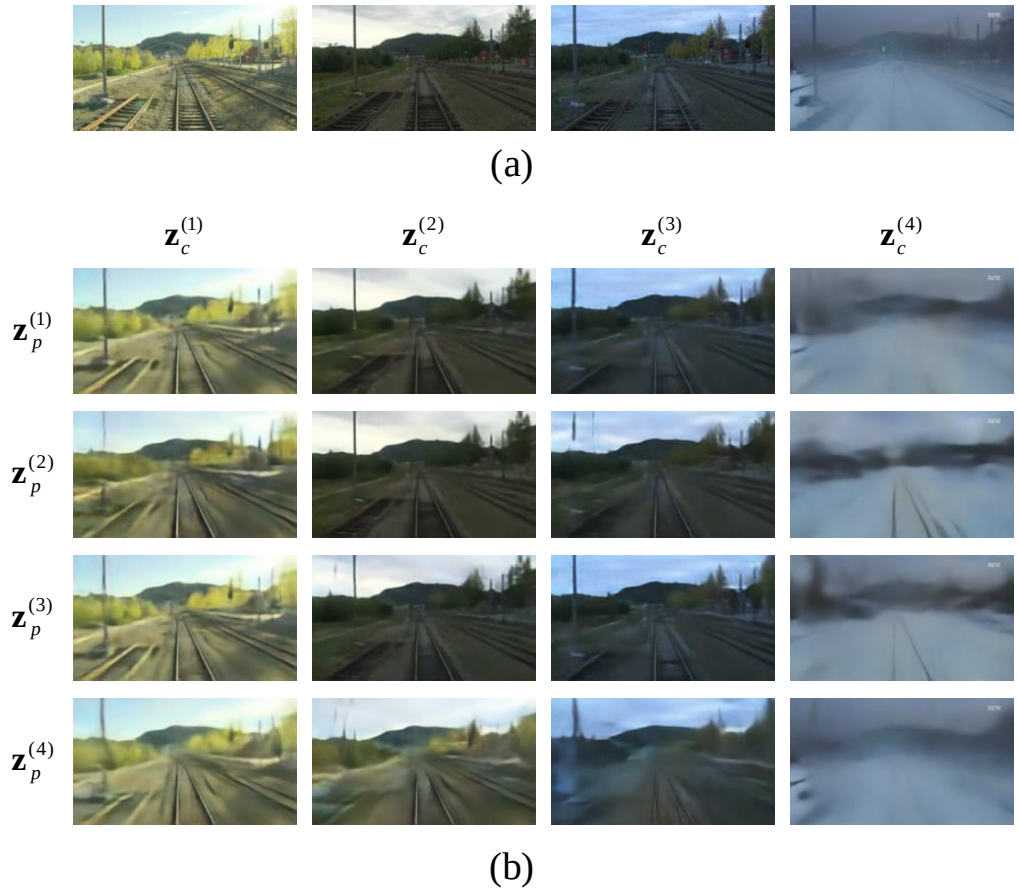

Figure 6. The independence visualization result of the Nordland dataset. (**a**) The first row is the original image, and (**b**) the other images are reconstructed by a combination of various $\mathbf{z}_p$ and $\mathbf{z}_c$.

As can be seen from the reconstructed image results in Figure 6, there is no significant change in the $\mathbf{z}_p$ change, whereas different season images are created in $\mathbf{z}_c$ change. Similarly, there is no significant difference in the change of $\mathbf{z}_p$, but it can be seen that images at different times are created according to the change of $\mathbf{z}_c$ in Figure 7. We can conclude that the environmental information is compressed in $\mathbf{z}_c$ since the reconstructed image is changed by the influence of the $\mathbf{z}_c$ rather than $\mathbf{z}_p$.

As the $\mathbf{z}_c$ plays a significant role in reconstructing the image, similar images would be generated if the same $\mathbf{z}_c$ is used to reconstruct the image. In other words, if we define ${}^{o}\mathbf{z}_c$ as a constant vector, $\{{}^{u}\mathbf{z}_p, {}^{o}\mathbf{z}_c\}$ and $\{{}^{v}\mathbf{z}_p, {}^{o}\mathbf{z}_c\}$ will reconstruct condition-invariant images ${}^{o}\mathbf{x}$ since the image is mainly affected by the condition-sensitive feature ${}^{o}\mathbf{z}_c$. The results of the condition-invariant image are shown in Figures 8 and 9.

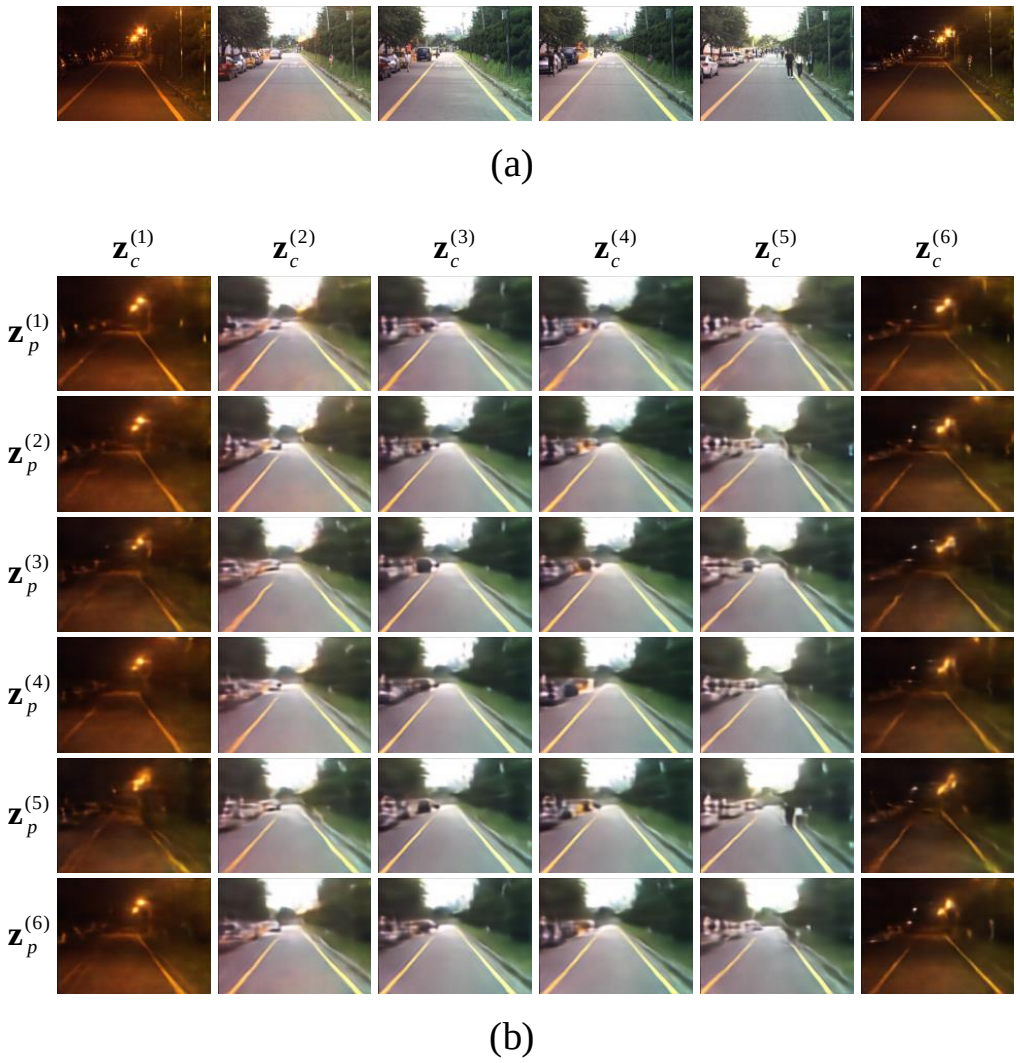

Figure 7. The independence visualization result of the Kaist dataset. (**a**) The first row is the original image, and (**b**) the other images are reconstructed by a combination of various $\mathbf{z}_p$ and $\mathbf{z}_c$.

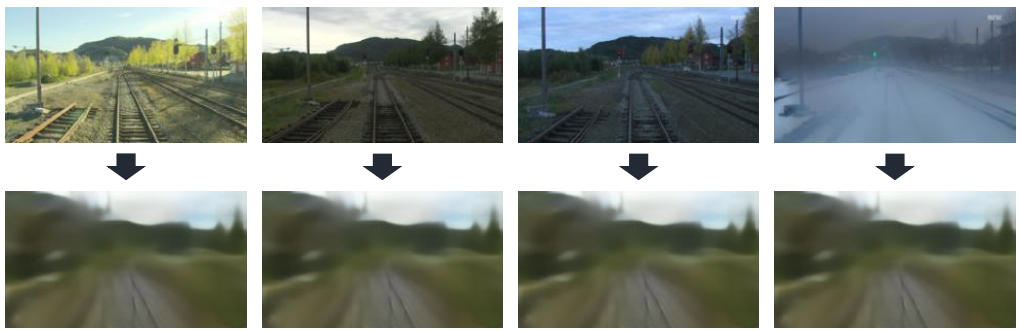

Figure 8. The condition-invariant image generation results using the constant feature vector $\mathbf{z}_c$ of the Nordland dataset.

Figure 9. The condition-invariant image generation results using the constant feature vector $\mathbf{z}_c$ of the Kaist dataset.

As expected, we can see that similar images are generated regardless of time or season changes if we use the same ${}^{o}\mathbf{z}_c$. The visualization results showed that the independent assumption between $\mathbf{z}_p$ and $\mathbf{z}_c$ is reasonable because the reconstructed images are mainly influenced by the condition-sensitive feature $\mathbf{z}_c$. Therefore, we can conclude that our model can extract the condition-invariant feature $\mathbf{z}_p$ and perform robust place recognition in changing environments using this feature.

To compare the place recognition performance, precision-recall analysis was conducted. Various thresholds were applied to the values of the similarity matrix. We compared the proposed method C-VAE (*VAE+C*) and CP-VAE (*VAE+C+P*) with the sum of the absolute difference (SAD) [24], FAB-MAP [25], AlexNet [10], and VGG19 [26]. The precision-recall results are shown in Figures 10 and 11.

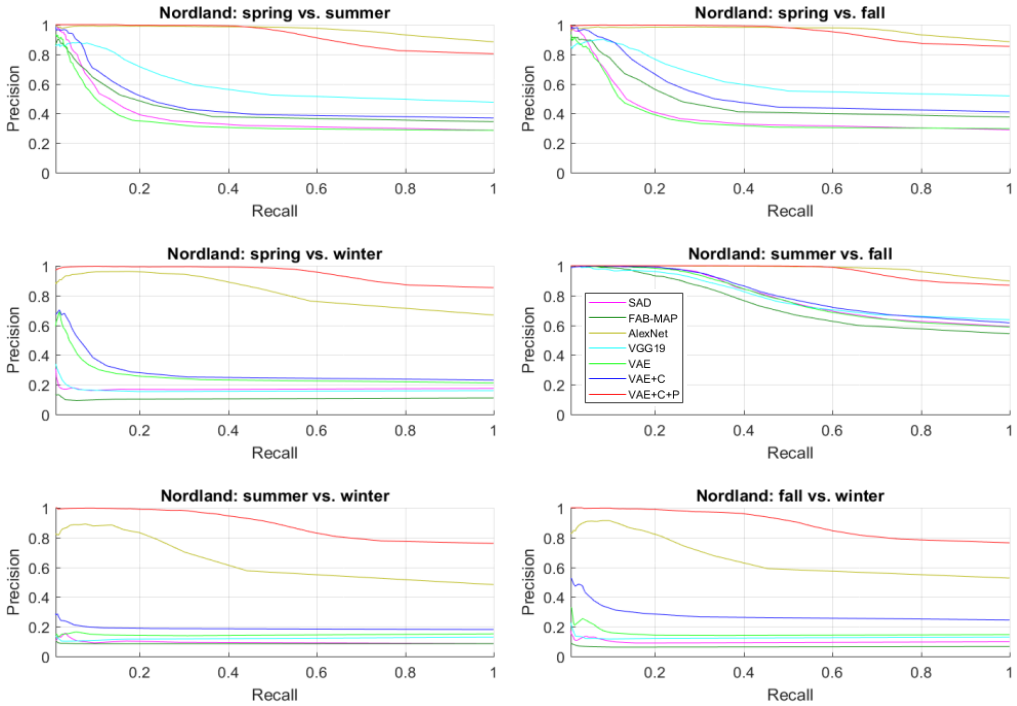

Figure 10. The precision-recall results in various seasons of the Nordland dataset.

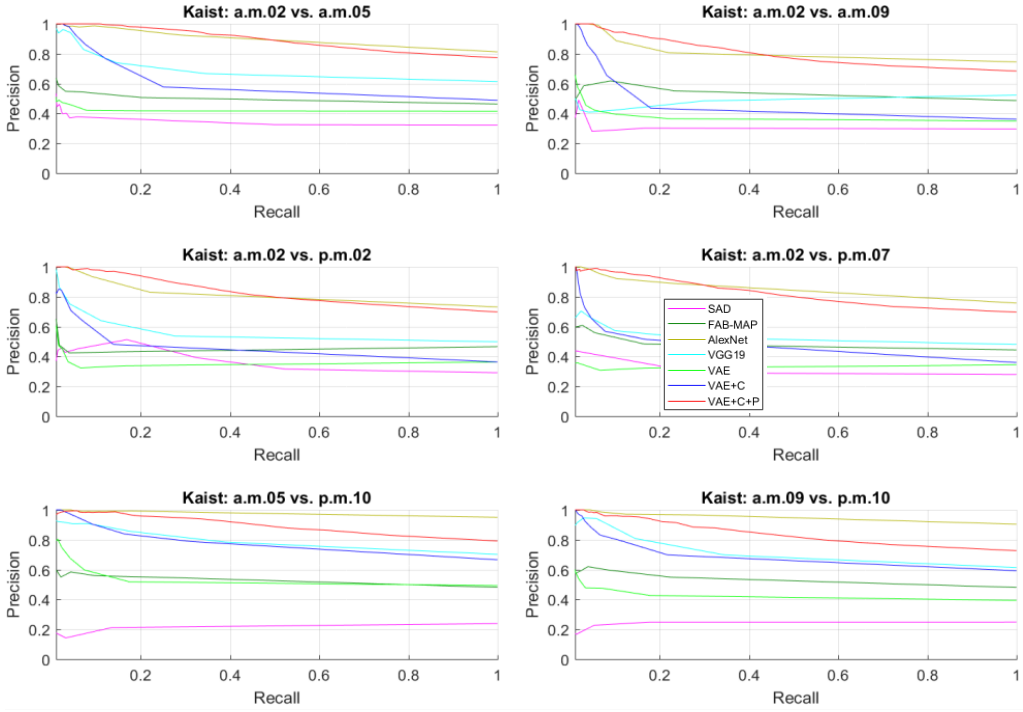

Figure 11. The precision-recall results in various seasons of the Kaist dataset.

Precision-recall results showed that the proposed method CP-VAE outperformed other methods in most cases. Existing handcraft features such as SAD and FAB–MAP showed they are not suitable for place recognition in a changing environment. Pre-trained deep learning models such as AlexNet and VGG19 showed reasonable performance in various situations. However, the performances were degraded when environmental changes between images were substantial, such as winter images with snow and other seasonal images without snow. This is a fatal weakness of the pre-trained model from the viewpoint of securing stability for long-term operation of the robot. Since the proposed method recognizes a place using condition-invariant features, it shows robustly high performance even in these cases. From the results of the precision-recall analysis, we were able to verify the validity of the proposed method's place recognition performances in a changing environment.

6. Conclusions

Variational Bayesian methods can perform efficient inference and learning in the presence of continuous latent variables with intractable posterior distributions, and large datasets. We introduced a stochastic variational inference and learning architecture that can extract condition-invariant features. Under the assumption that a latent representation of the variational autoencoder can be divided into condition-invariant and condition-sensitive features, a new structure of the variation autoencoder is proposed and a variational lower bound is derived to train the model. After training the model, condition-invariant features are extracted from test images to calculate the similarity between them, and the places can be recognized even in severe environmental changes. Experimental results showed that our assumption was reasonable, and the validity of the proposed method was proved by the precision-recall analysis. In the future, it is necessary to develop a method that can be applied even when several environmental factors are mixed. For example, if we develop a place recognition method that is robust to seasonal and weather changes, the robot will be able to operate in a variety of environmental conditions.

Author Contributions: Conceptualization, J.O.; methodology, J.O.; validation, J.O.; investigation, J.O. and G.E.; writing—original draft preparation, J.O.; writing—review and editing, J.O.; project administration, J.O.; funding acquisition, J.O. Both authors have read and agreed to the published version of the manuscript.

Funding: This work has supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2020R1F1A1076667), Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20174010201620). This work was also supported by Research Resettlement Fund for the new faculty of Kwangwoon University in 2019.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

SLAM	Simultaneous Localization And Mapping
SIFT	Scale Invariant Feature Transform
SURF	Speeded Up Robust Features
HOG	Histogram of Oriented Gradients
CNNs	Convolutional Neural Networks
CAEs	Convolutional Auto Encoders
VAEs	Variational Auto Encoders

References

1. Lowry, S.; Sünderhauf, 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]
2. Sattler, T.; Maddern, W.; Toft, C.; Torii, A.; Hammarstrand, L.; Stenborg, E.; Safari, D.; Okutomi, M.; Pollefeys, M.; Sivic, J.; et al. 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, 18–22 June 2018; pp. 8601–8610.
3. Sünderhauf, N.; Protzel, P. BRIEF-Gist—Closing the loop by simple means. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Brisbane, Australia, 25–30 September 2011; pp. 1234–1241. [CrossRef]
4. Liu, Y.; Zhang, H. Visual loop closure detection with a compact image descriptor. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura-Algarve, Portugal, 7–12 October 2012; pp. 1051–1056.
5. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
6. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). *Comput. Vis. Image. Und.* **2008**, *110*, 346–359. [CrossRef]
7. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
8. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
9. Torralba, A.; Fergus, R.; Weiss, Y. Small codes and large image databases for recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
10. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. In Proceedings of the International Conference on Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
11. Arandjelović, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1437–1451. [CrossRef]
12. Chancán, M.; Hernandez-Nunez, L.; Narendra, A.; Barron, A.B.; Milford, M. A hybrid compact neural architecture for visual place recognition. *IEEE Robot. Autom. Lett.* **2020**, *5*, 993–1000. [CrossRef]
13. Sünderhauf, N.; Shirazi, S.; Jacobson, A.; Dayoub, F.; Pepperell, E.; Upcroft, B.; Milford, M. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. In Proceedings of the International Conference on Robotics: Science and Systems. Robotics: Science and Systems Conference, Rome, Italy, 13–17 July 2015; pp. 1–10.

14. Garg, S.; Sünderhauf, N.; Milford, M. Don't look back: Robustifying place categorization for viewpoint-and condition-invariant place recognition. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–26 May 2018; pp. 3645–3652.
15. Naseer, T.; Ruhnke, M.; Stachniss, C.; Spinello, L.; Burgard, W. Robust visual SLAM across seasons. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 2529–2535.
16. Oh, J.H.; Lee, B.H. Dynamic programming approach to visual place recognition in changing environments. *Electron. Lett.* **2017**, *53*, 391–393. [CrossRef]
17. Park, C.; Chae, H.W.; Song, J.B. Robust Place Recognition Using Illumination-compensated Image-based Deep Convolutional Autoencoder Features. *Int. J. Control Autom. Syst.* **2020**, *18*, 2699–2707. [CrossRef]
18. Kingma, D.; Welling, M. Auto-encoding variational Bayes. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
19. Pu, Y.; Gan, Z.; Henao, R.; Yuan, X.; Li, C.; Stevens, A.; Carin, L. Variational autoencoder for deep learning of images, labels and captions. In Proceedings of the International Conferene on Advances in neural information processing systems (NIPS), Barcelona, Spain, 5–10 December 2016; Volume 29, pp. 2352–2360.
20. Oh, J.; Han, C.; Lee, S. Condition-invariant robot localization using global sequence alignment of deep features. *Sensors* **2021**, *21*, 4103. [CrossRef] [PubMed]
21. Sünderhauf, N.; Neubert, P.; Protzel, P. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In Proceedings of the Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA), Karlsruhe, Germany, 6–10 May 2013.
22. Olid, D.; Fácil, J.M.; Civera, J. Single-view place recognition under seasonal changes. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Workshops, Madrid, Spain, 1–5 October 2018.
23. Choi, Y.; Kim, N.; Park, K.; Hwang, S.; Yoon, J.; Kweon, I.S. All-day visual place recognition: Benchmark dataset and baseline. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Boston, MA, USA, 7–12 June 2015.
24. Milford, M.; Wyeth, G. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), St Paul, MN, USA, 14–19 May 2012; pp. 1643–1649.
25. Cummins, M.; Newman, P. Appearance-only SLAM at large scale with FAB-MAP 2.0. *Int. J. Robot. Res.* **2011**, *30*, 1100–1123. [CrossRef]
26. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2015.