# Vehicle localization based on scene perception

Linrunjia Liu

**THÈSE DE DOCTORAT DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ**

**PRÉPARÉE À L'UNIVERSITÉ DE TECHNOLOGIE DE BELFORT-MONTBÉLIARD**

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Informatique

par

LINRUNJIA LIU

# Vehicle Localization Based on Scene Perception

Thèse présentée et soutenue à Belfort, le 30 Juin 2021

Composition du Jury :

| | | |
|---|---|---|
| TALEB-AHMED ABDELMALIK | Professeur à l'Université Polytechnique des Hauts de France | Rapporteur |
| DORNAIKA FADI | Professeur à Université du Pays Basque | Rapporteur |
| NOYER JEAN-CHARLES | Professeur à l'Université du Littoral Cote d'Opale | Examinateur |
| MEURIE CYRIL | Chargé de recherche à Université Gustave Eiffel | Examinateur |
| CAPPELLE CINDY | Maître de conférences à l'Université Bourgogne Franche - Comté, UTBM | Co-encadrante |
| RUICHEK YASSINE | Professeur à l'Université Bourgogne Franche - Comté, UTBM | Directeur de thèse |

N° | X | X | X |

ÉPIM
école doctorale **sciences pour l'ingénieur et microtechniques**

**Title:** Vehicle Localization Based on Scene Perception

**Keywords:** Place recognition, Feature representation, Dynamic objects removal, Image blurring, Image enhancement

**Abstract:**

The vision-based vehicle localization task is tackled as an image based place recognition problem in this thesis. As image representation is a crucial step in place recognition process, we proposed place recognition methods focusing on increasing the quality of image features. First, dynamic object removal step is proposed to remove the dynamic objects, such as vehicles and pedestrians, from the images by semantic segmentation method and then background information restoring by image inpainting method. Second, reducing instead of removing the noisy information of an image is proposed. Using image blurring method to reduce the image noise, the performance of place recognition is improved with less time consuming compared to the dynamic objects removal step. Third, feature similarity achievement method is proposed for the day and night place recognition problem. Blurring the daytime images and enhancing the nighttime images to achieve a balance in their feature quality is the key to improve the place recognition performance. The images generated by the proposed methods are evaluated in the same place recognition framework and proved their effectiveness by comparison with other advanced place recognition methods.

**Titre :** Vehicle Localization Based on Scene Perception

**Mots-clés :** Reconnaissance de lieu, Représentation d'images, Suppression d'objets dynamiques, Flou d'image, Amélioration d'image

**Résumé :**

Dans cette thèse, la tache de localisation de véhicule basée sur la vision est abordée comme un problème de reconnaissance de lieu. La représentation de l'image étant une étape essentielle dans le processus de reconnaissance de lieux, donc des méthodes de reconnaissance de lieux axées sur l'augmentation de la qualité des caractéristiques de l'image sont proposées. Tout d'abord, une étape de detéction et suppression d'objets dynamiques tels que des véhicules et des piétons est proposée. Les objets dynamiques sont détectés par segmentation sémantique puis supprimés, les parties manquantes sont alors reconstruites par une approache d'inpainting (restauration) de l'arrière-plan. Dans un deuxième temps, il est proposé de réduire au lieu de supprimer de l'image les informations non permanentes de la scène (les objets dynamiques). En appliquant un flou sur l'image pour réduire le bruit de l'image, les performances de la reconnaissance de lieu sont améliorées avec un coût calculatoire moindre que l'étape de suppression d'objets dynamiques et reconstruction du fond. Dans un troisième temps, on s'intéresse à la reconnaissance de lieu dans le cas particulier et difficile jour/huit. Pour cela, l'idée est d'équilibrer la qualité des images entre les conditions jour et nuit. Les images diurnes (souvent de meilleures qualité) sont alors dégradées alors que les images nocturnes sont améliorées.

L'apport des images générées par ces différentes méthodes de pré-traitement proposées est évalué dans le cadre de la même approche de reconnaissance de lieu et leur efficacité est prouvée par comparaison avec d'autres méthodes avancées de reconnaissance de lieu de la littérature.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ
Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon, France

# ACKNOWLEDGEMENTS

I would like to particularly thank my supervisors, Prof. Yassine Ruichek and Prof. Cindy Cappelle. This research could not have been accomplished without their sustaining support, guidance, patience and encouragement all the way.

I would like to acknowledge my thesis committee, Abdelmalik Taleb-Ahmed, Fadi Dornaika, Jean-Charles Noyer and Cyril Meurie, for accepting and evaluating my thesis work.

I am also grateful to the China Scholarship Council (CSC) for putting faith in my studies and providing the financial support for this research at the University of Bourgogne Franche - Comté, UTBM in France.

Besides, I would like sincerely to thank a number of colleagues and friends: Shuang, Yue, Kui, Shengrong, Hongjian, Weizhu, Yang, Shiming, Tao, Rong, Citlali, Karen, Hui, Hongyuan, Wenheng, Ziqing, Rui, Gaoshuai, Ao and Shuohong, for providing a welcoming and supportive community through this memorable journey.

In the end, a big big kiss to my beloved parents, sister and boyfriend. Their love gave me the courage to pass through the difficult time that appeared frequently during the research process.

# CONTENTS

# I

# Context and Problems

# 1

# INTRODUCTION

This thesis aims to tackle the environment perception based vehicle localization problem, which is a fundamental task in self-driving cars. The environment perception part can be achieve by the sensors of the intelligent vehicle like monocular camera and the localization part can be achieve by GPS information. The principle of the perception based localization concept followed in this thesis is to find the image from a priori reference images that best matches with the current captured image. It deals with then an image based place recognition problem. Approaches are proposed in this thesis to improve the image based place recognition performance by increasing the effectiveness of image features.

This chapter helps to have a better understanding of the topic of this thesis by introducing the background and explaining the problem and objectives. The framework of the Automated Driving Systems (ADS) is described in Section 1.1. Section 1.2 details the process of scene perception based vehicle localization task and points out that the core part is place recognition. Then the problem and difficulty of place recognition are stated in Section 1.3. Besides, the objectives and contribution works are also exhibited in this section. Finally, the organization of this thesis is given at the end of this chapter.

## 1.1/ BACKGROUND

Since 94% of road accidents are resulting from human errors according to the report of National Highway Traffic Safety Administration (NHTSA) [Singh, 2015], the development of ADS promises a safer driving environment. One of the earliest research project about self-driving vehicles (Eureka Project PROMETHEUS) [Eureka, ] was developed

from 1987 to 1995 in Europe. Based on it, VITA II [Ulmer, 1994] by Daimler-Benz was then carried out.

According to the definition of the Society of Automotive Engineers (SAE) [SAE, 2018], the automation level of automobiles is as follows:

- L0: No automation;
- L1: Primitive driver assistance system, including adaptive cruise control, anti-lock braking systems [Rajamani, 2006], etc ;
- L2: Partially automated. The advanced assistance system includes emergency braking or collision avoidance [Hafner et al., 2013] [Colombo et al., 2012];
- L3: Conditional automation. The vehicle can achieve self-driving during normal operation. But once an emergency takes place, the driver should focus on driving immediately;
- L4: When the weather conditions permit and the infrastructure (network, maps, etc) is complete, no driver is required;
- L5: No driver is needed in any case.

There is no produced vehicle that achieves L4 level and above in driving automation at present. The environment complexity, such as changing weather condition and unpredictable human behavior, makes the vehicle difficult to recognize the situation and decide its reaction. For example, because of the failure estimation of the speed of a bus, Google self-driving car crashed into the bus. The Autopilot of Tesla hit a white truck and killed the driver because it failed to recognize the truck. The development of ADS has overcame some of the previous problems but there is still a long way to go to achieve safely automated driving under strict conditions.

According to the work of [Yurtsever et al., 2020], the ADS high-level system architecture is classified as in the Fig. 1.1 based on the connectivity and algorithm implementation. For the connectivity, the framework of ADS can be divided into Ego-only systems [Levinson et al., 2011] and Connected multi-agent systems [Gerla et al., 2014] [Lee et al., 2016] [Amadeo et al., 2016]; in terms of algorithm implementation, it can be divided into two categories, one is by modularizing each part to achieve

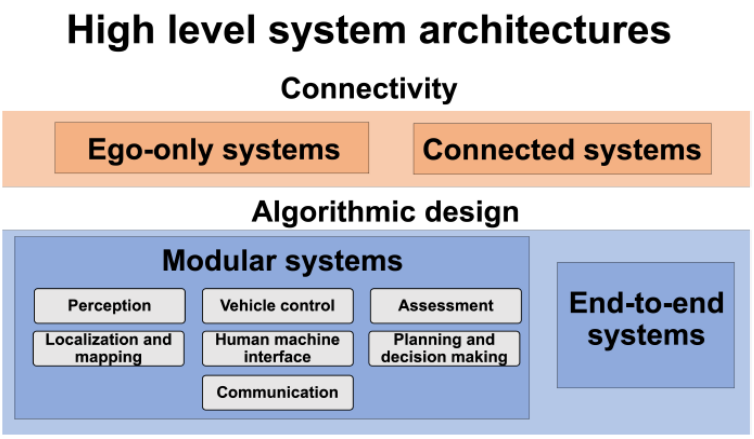



Figure 1.1: A high level classification of automated driving system architectures [Yurtsever et al., 2020].

[Levinson et al., 2011] [Wei et al., 2013] [Broggi et al., 2013], the other is through the end-to-end system to achieve [Chen et al., 2015a] [Baluja, 1996] [Koutník et al., 2013]. The core functions of the modular system include perception, localization, mapping, planning and human machine interaction.

This thesis aims to contribute to the localization function and focuses on vehicle localization based on scene perception.

## 1.2/ VEHICLE LOCALIZATION BASED ON SCENE PERCEPTION

Localization is a fundamental task for self-driving cars, it permits the vehicle to estimate its position in its environment. The three common approaches of vehicle localization are [Yurtsever et al., 2020]: GPS-IMU fusion, simultaneous localization and mapping (SLAM), and a priori map-based localization.

In the GPS-IMU fusion approach [Urmson et al., 2004] [Zhang et al., 2012], IMU is used to estimate the dead-reckoning position of the vehicle while GPS provides the absolute position data. When IMU locates the vehicle by dead reckoning, errors will accumulate with time. By being fused with GPS data, the errors of IMU will be corrected, thus avoiding the failure of long-term navigation.

The accuracy of the GPS-IMU fusion method is relatively low, and in fact cannot be di-

rectly used for vehicle positioning. In addition, in dense urban environments, the accuracy of GPS will be affected by tunnels, high-rise buildings, etc. Even though, combining the GPS-IMU system with sensors such as lidar for pose estimation can achieve better performance.

SLAM [Bresson et al., 2017] is the method that finds the position of the vehicle in an environment and at the same time makes the map of this environment online. In theory, SLAM does not need a priori information about the environment. Therefore, it can be applied in any environment. However, it is more used in indoor environments.

In outdoor environments, localization is more often based on an apriori map. The key point of the apriori map-based localization is matching. By comparing the image obtained in the current position with images in the apriori map, the location of the best matching image can be found, achieving thus the localization task.

The task of this thesis is to achieve the vehicle localization based on scene perception, which is an apriori map-based localization process, as seen in Fig.1.2.

Perceiving the environment and getting the information is the first step for localization. Monocular camera is one of the most common and cheapest sensor to get a 2D image of the current environment. The current image is seen as the query image and will be compared with each GPS-tagged image in the apriori map called reference in the following. By feature matching, the best matching image will be found. Finally, the localization will be achieved by obtaining the GPS information of the reference image that best matches with the current query image.

The biggest drawback of the apriori map-based localization method is the acquisition of an apriori map. In fact, making and maintaining a reliable high-precision map is quite time-consuming and laborious. Therefore, instead of making the apriori map, this thesis directly uses the GPS-tagged apriori map made by other researches and concerns more about the matching process of the query and reference images, which is in fact a place recognition process. Once the place recognition result is obtained, the location information can be read by the tagged GPS data of the best matching image.

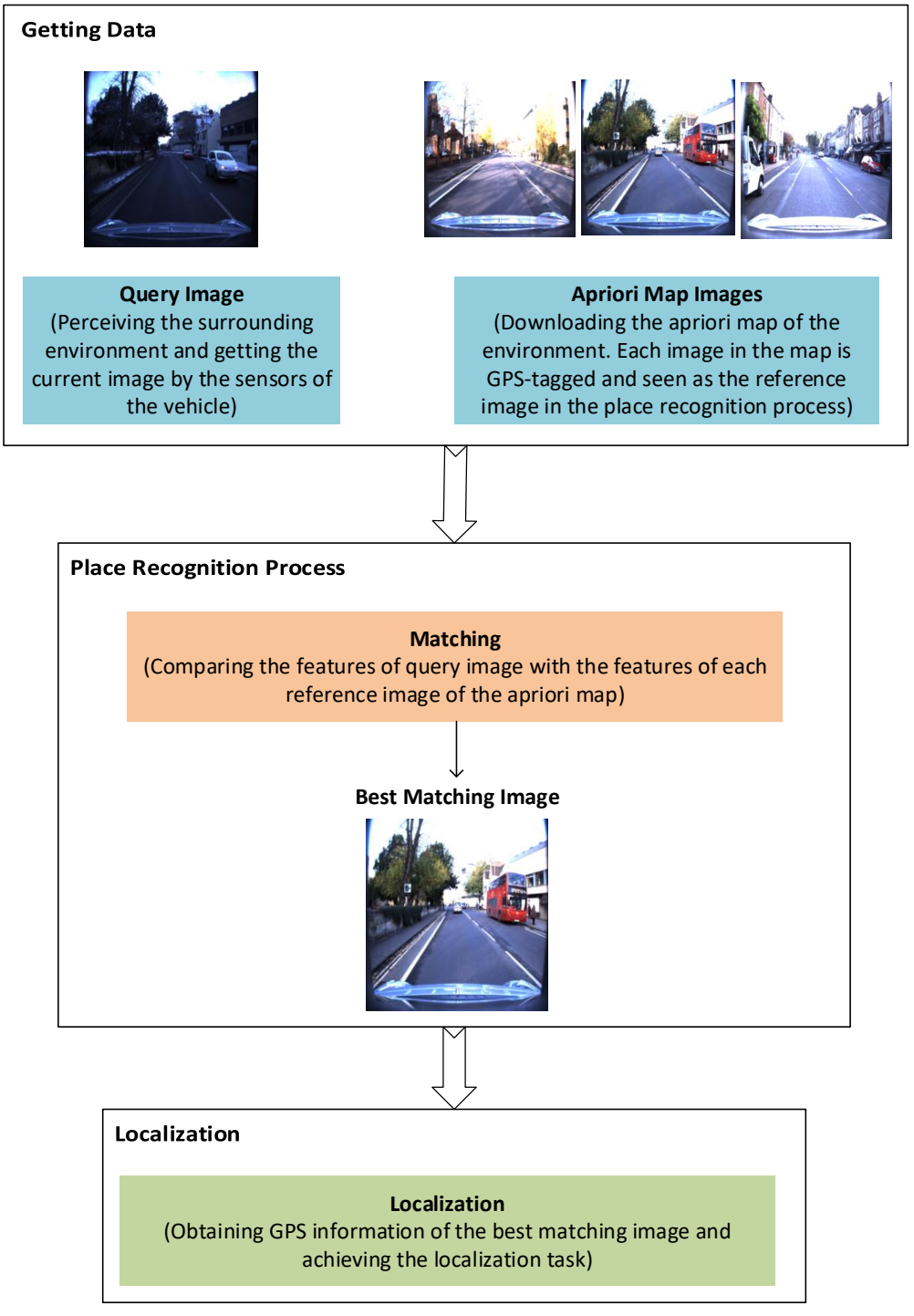



Figure 1.2: The process of vehicle localization based on scene perception.

## 1.3/ Problem Statement and Objectives

Place recognition for vehicle localization aims to select the image in a long-term robot-car dataset, which represents the same place as a query image. Place recognition has progressed significantly with the development of deep learning. Nevertheless, due to the severe appearance changes in long-term robot navigation, there are still many challenges and problems to achieve efficient image based place recognition. The appearance changes caused by dramatic changes in weather conditions, illumination, viewpoint, and even dynamic objects, such as vehicles and pedestrians, make it difficult to distinguish the same place at different times. This thesis focuses on the appearance variations and proposes methods to decrease the appearance difference of the query and reference images. Without the noisy information, features generated from each image are more effective in the place recognition process.

The three main contributions of this thesis are detailed as follows:

**1)** The first contribution is to remove dynamic objects from images. Inspiring by the state-of-the-art performance of place recognition that applies invariant features of an image, our approach tries to delete all the dynamic objects in an image and extract features by the remaining stable background. The development of semantic segmentation and image inpainting approaches make it possible. Using semantic segmentation method, dynamic objects, objects classified as cars, buses, trucks, and pedestrians, can be well recognized and masked. Using image inpainting method, the labelled dynamic objects in the raw images will be seen as missing parts and be filled by fine details of the background information. Relative to the raw images, the inpainted images without the dynamic objects have better performance in place recognition according to the experiment results.

**2)** When the inpainted images are used for place recognition in our first contribution work, the results show that the place recognition performance can be improved even if the dynamic objects are not perfectly removed and the background information is restored with some noise. Therefore, reducing instead of removing the noisy information is a direction worth to try as it is much easier to achieve. By applying image blurring method to the reference sequence which is often acquired in better condition than the query sequence, the features of the two sequences can achieve

a balance and improve the place recognition performance. Recognizing the images with dynamic objects and blurring them is then the second method exploited in this thesis to reduce the noise in the images. This method can prove its effectiveness in the busy traffic conditions.

**3)** The day and night place recognition problem is hard to solve due to the illumination and appearance changes. By translating the nighttime images into daytime ones using image-to-image translation methods, researches prove that having a similar appearance of query and reference images can indeed help to improve the place recognition performance. However, facing the low-quality of nighttime image, the translated daytime image obtained by the image translation method has poor quality and is different from the reference daytime image. To tackle this, the day and night place recognition is divided into two categories: night vs. day and day vs. night place recognition.

For night vs. day place recognition, we demonstrate instead of translating the nighttime images into daytime ones, enhancing the nighttime query image as well as weakening the daytime reference image can also achieve feature similarity of query and reference images and improve the place recognition performance. This new feature similarity achievement for day and night place recognition approach is developed by combining: 1) nighttime image enhancement method, and 2) daytime image blurring method. Besides, a dynamic objects removing process is also proposed and combined with a stable feature selection method.

For day vs. night place recognition, the query day image has good feature quality while the reference night images are captured under bad condition. Methods are used to enhance the effectiveness of nighttime image features. The hard light regions are repaired by image inpainting method and under-exposed regions are enhanced by image enhancement method.

## 1.4/ THESIS ORGANIZATION

The rest of the thesis is organized as follows.

**Chapter 2**: To have a knowledge of the development of image perception based vehicle

localization, related researches are reviewed in this chapter. Feature representation and matching are two important steps for image perception based localization. According to the feature representation methods, the localization researches can be divided into three categories: approaches based on hand-crafted features which include global features and local features, approaches based on CNN features and approaches based on synthetic features. According to the feature matching methods, the way to tackle localization problem based on global feature matching methods and region or landmark based feature matching methods are reviewed.

**Chapter 3**: As the place recognition framework and experimental setup of our three contributions are the same, this chapter gives an overview of the place recognition system, the evaluation dataset as well as the evaluation metric considered in this thesis.

**Chapter 4**: Improving place recognition by removing dynamic objects using inpainted images is proposed in this chapter. To have a better understanding, semantic segmentation and image inpainting are explained firstly. Then the process of generating the proposed inpainted images is exhibited in details. Two different methods are used to restore the background information of the removed dynamic objects of images: single image inpainting method and sequence image inpainting method. In the experimental part, the performance of the proposed methods is evaluated in different aspects. Place recognition results are obtained based on four different CNN models, and with extended ten sequences to evaluate the performance of the use of inpainted images. Besides, the influence of using both the inpainted query and reference sequences or using only one of them is also studied.

**Chapter 5**: Reducing the image noise by image blurring method is proposed in this chapter. By applying Gaussian blurring to the image, the whole blurred-image is generated. Inspired by the removing dynamic objects method introduced in Chapter 4, the dynamic objects of images recognized by semantic segmentation method are blurred in this chapter, thus generating the blurred-fcn image. The efficiency of these blurred images are evaluated under all the daytime conditions, day vs. day, day vs. night and night vs. night. The results show that the method of reducing noise is very simple yet effective to improve the place recognition performance. Same as Chapter 4, only blurring the reference image is considered (offline process) and can assure good performance of place recognition in most conditions. The performance of the proposed method is compared with two

advanced methods: DenseVLAD [Torii et al., 2015] and OLO [Chen et al., 2017b].

**Chapter 6**: Methods to tackle the day and night place recognition are proposed in this chapter. This problem can be divided into night vs. day and day vs. night place recognition. The difference of these two problems is in the quality of the query image . For the night vs. day place recognition task, the query image has bad quality, so enhancing the query nighttime images and blurring the daytime reference images can balance the quality of image features and improve the place recognition performance. For the day vs. night place recognition, decreasing the quality of the query day image which is good in general (compared to nighttime) will lead to the failure of place recognition task. Therefore, the key point is to increase the nighttime reference images to achieve the good place recognition performance. Combining the image enhancing method with the removing dynamic objects proposed in Chapter 4 is a good way to increase the information of the low-exposed regions and decrease the dynamic information of nighttime images. The results of the proposed methods are compared with DenseVLAD and the night to day translation method, ToDayGAN.

**Chapter 7**: To summarize the place recognition methods proposed in this thesis and plan the future work, a conclusion is made in this chapter.

# 2

# Related Works of scene perception based vehicle localization

This chapter presents the related researches of scene perception based vehicle localization. It begins by addressing the localization problem. Then the core aspects to tackle this problem is pointed out: image representation and feature matching. The scene perception based localization approaches based on image representation and feature matching are further reviewed in this chapter.

## 2.1/ Introduction

Scene perception based vehicle localization aims to find the pose (position and orientation) of the image taken by car sensors. This is an image-based localization problem, which can be also called as visual localization or visual-based localization problem as there is no standardized designation.

As vehicle localization is a well researched topic, methods to tackle this problem are various. According to [Piasco et al., 2018], image-based localization can be classified as indirect methods [Arandjelović et al., 2012] [Radenović et al., 2016] as well as direct methods [Kendall et al., 2015] [Sattler et al., 2017], as illustrated in Fig. 2.1. The indirect methods get a coarse pose information by finding the best matching image corresponding to the query image and consider its GPS information as the location of the


Figure 2.1: Examples of image-based localization systems [Piasco et al., 2018].(a) Indirect method from [Radenović et al., 2016] retrieved top-k ranking similar images to the query image. (b) Direct method from [Feng et al., 2016] shows that the features of two query images are aligned with 3D points, respectively.

query image. This can be seen as the place recognition problem, as detailed in Chapter 1. The direct methods recover instantly the exact 6DoF (Degrees of Freedom) pose of the query image which is more accurate than the coarse estimation recovered from the indirect method. In the direct method, a known image acquisition system, such as SLAM [Milford et al., 2012], augmented reality [Arth et al., 2015] or visual odometry [Pascoe et al., 2015] is needed to get the relative pose of database. Then, as shown in Fig. 2.1 (b), the 2D features of the query image are matched with the 3D points constructed offline to estimate the camera pose of the query image.

This thesis focus on the indirect method of the image-based localization and divides this problem into four modules, as seen in Fig. 2.2.

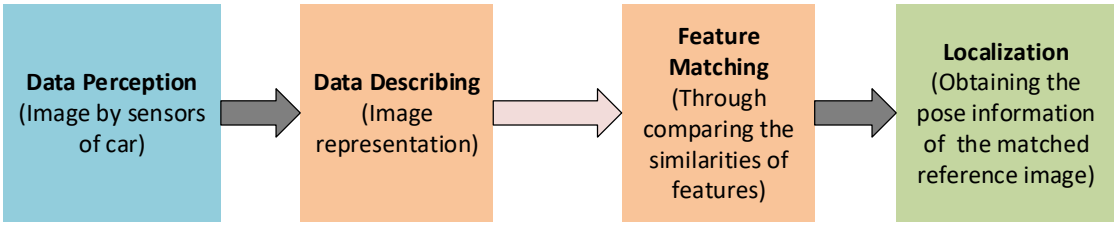



Figure 2.2: The main modules of image perception based localization.

- Data perception. Obtaining the data of the surrounding environment is a fundamental task for localization. The types of the data can vary from image (captured by monocular camera in common), geometric information (captured by RGB-D camera, LIDAR) to semantic clues. Combining the different data can provide more

possibilities to tackle the difficulties of localization.

- Data Describing. Instead of directly using them, data needs to be described in order to be better stored and recognized. For image data, traditional representation methods used to represent images are the hand-crafted features. With the development of deep learning, the image representations have evolved to deep-learned features. Besides, image representations can also be divided by extracting the features of whole the image or selecting the interested regions where to extract.

- Feature Matching. This is the retrieval process. By comparing the features of the query image with images in the apriori map (reference images), the best image matching will be found. Distance computation is a very simple method and is often used to match the feature vectors. When the number of features is large, nearest neighbour search is a good alternative for the similarity research. Besides, machine learning methods, such as SVM classifier, Multi-Task Learning and Gaussian Mixture Model are other choices of feature comparison.

- Localization. The final step for the image-based localization is to get the localization information form the matched reference image. The localization information has been labelled offline to each image in the apriori map with for example GNSS (as GPS) based measurement. As the localization result is the GPS data of the best matched image (which is not necessary at the exact same location as the query image), this is a coarse localization process and can be used as an initial information for further fine localization.

Though seen as the place recognition process, image perception based vehicle localization focuses only on the environment around the route. This is an outdoor environment whose appearance is strongly influenced by illumination, weather, season, time over a day, traffic condition, etc. Facing to the severe changing of the scene appearance, the data describing module to represent the image features and the feature matching module play important roles in the localization performance. Therefore, the related works of the image perception based vehicle localization in this thesis are divided into two categories: approaches based on image representation and approaches based on feature matching.

In this chapter, the approaches of image-based localization as well as approaches of place recognition are mainly reviewed. The rest of this chapter is divided as follows: Sec-

tion 2.2 introduces the related work based on image representation method, which gives an overview for the development of the feature descriptors used in the image-based localization, from hand-crafted features, deep-learned features to synthetic features. Section 2.3 describes the related work based on the feature matching method, which includes the nearest neighbour searching and machine learning based matching.

## 2.2/ Approaches for Image Representation stage

Image representation is the basis step for visual place localization. Since extracting robust features determines the performance of visual place recognition, there are lots of visual place localization works focusing on image representation methods. From traditional hand crafted features to novel translated features obtained by Generative Adversarial Networks (GANs), the feature extraction methods benefit a lot from the development of computer vision technology.

### 2.2.1/ Approaches based on hand-crafted features

Local feature methods such as scale-invariant feature transforms (SIFT) [Lowe, 2004] and speeded-up robust features (SURF) [Bay et al., 2006a], as well as global feature methods such as Gist [Oliva et al., 2001] have been applied to visual place recognition in the traditional research work before the advent of CNNs. The difference of local features and global features is that local features describe only Regions of Interest (ROI) of an image chosen by different measures, while global features have not a selection process for an image and describe the image as a whole. The following discussion provides the understanding of image representation based on not only local features but also global ones.

#### 2.2.1.1/ Local features

Local features describe specific regions of an image. These regions can be sparsely described, in a two steps process: (1) feature detection, (2) feature description. In the feature detection step, the salient structure of an image, such as corner, blob or region is detected and represented as a key point. Table 2.1 shows some famous detectors with

the introduction of their types and references: FAST, ORB, BRISK, SIFT, SURF, Star and MESR. Then the key points are characterized according to the neighbourhood and can be described as a multi-dimensional vector. The descriptors with their dimensions and types are shown in Table2.1: SIFT, SURF, ASIFT, BRIEF, ORB, BRISK, FREAK, D-BRIEF. These are the main local features used in localization, though it is not a complete summary. A comprehensive summary of local features for vision based topological mapping and localization can be found in the work of [Garcia-Fidalgo et al., 2015].

Table 2.1: Summary of some local features.

| Name | Feature | Type | Dimensions | Reference |
|---|---|---|---|---|
| FAST | detector | Corners | - | [Rosten et al., 2006] |
| ORB | detector | Corners | - | [Rublee et al., 2011] |
| BRISK | detector | Corners | - | [Leutenegger et al., 2011] |
| SIFT | detector | Blobs | - | [Lowe, 2004] |
| SURF | detector | Blobs | - | [Bay et al., 2006a] |
| Star | detector | Blobs | - | [Konolige et al., 2010] |
| MESR | detector | Regions | - | [Matas et al., 2004] |
| SIFT | descriptor | Float | 128 | [Lowe, 2004] |
| SURF | descriptor | Float | 32, 64, 128 | [Bay et al., 2006a] |
| ASIFT | descriptor | Float | 128 | [Morel et al., 2009] |
| BRIEF | descriptor | Bit | 128, 256, 512 | [Calonder et al., 2010] |
| ORB | descriptor | Bit | 256 | [Rublee et al., 2011] |
| BRISK | descriptor | Bit | 512 | [Leutenegger et al., 2011] |
| FREAK | descriptor | Bit | 512 | [Alahi et al., 2012] |
| D-BRIEF | descriptor | Bit | 32 | [Trzcinski et al., 2012] |

Table 2.2 lists some localization approaches based on local features. The most common used local features in image-based localization are SIFT, and SURF, as shown in Table 2.2. The SIFT feature is popular because of its invariance to rotation, scale scaling and brightness changes. Besides, it contains plenty of information and has good distinctiveness, making it suitable for fast and accurate matching. The SIFT algorithm aims to find key points in different scale spaces, which is achieved by four stages: 1) scale-space extrema detection, 2) keypoint localization, 3) orientation assignment, and 4) keypoint

Table 2.2: Summary of some image-based localization approach based on local features.

| References | Camera | Tasks | Environment | Feature |
|---|---|---|---|---|
| [He et al., 2006] | Mono | Map + Loc | Outdoor | SIFT |
| [Valgren et al., 2007] | Omnidir | Loc | Outdoor | SIFT/SURF |
| [Ascani et al., 2008] | Omnidir | Loc | In + Out | SIFT/SURF |
| [Anati et al., 2009] | Omnidir | Map + Loc | In + Out | SIFT |
| [Booij et al., 2009] | Omnidir | Map + Loc | In + Out | SIFT |
| [Badino et al., 2011] | Mono | Map + Loc | Outdoor | SURF/U-SURF |
| [Romero et al., 2010] | Omnidir | SLAM | Outdoor | MSER |
| [Majdik et al., 2013] | Mono | Loc | Outdoor | ASIFT |

descriptor. Though robust, the speed of SIFT algorithm is relatively slow, so a new algorithm, SURF, was firstly introduced in [Bay et al., 2006a] as an accelerated version of SIFT. SURF can achieve almost real-time matching of the objects in two images under moderate conditions. This is in fact based on the haar derivation to the integral image. Without considering the rotation invariance, the efficiency of the SURF algorithm is improved further more, thus Upright SURF (U-SURF) is proposed. Even without rotation invariance, U-SURF can have better performance than SURF in the task of long-term outdoor localization, what is proved in the work of [Valgren et al., 2007]. Beside the SIFT and SURF features, the localization approach in [Romero et al., 2010] uses the MSER feature to detect the blobs in the image and then extract them into vector. In this approach each image is divided into regions firstly, then the MSER features of each region can be grouped in a graph. Finally, by applying the method of Graph Transformation Matching (GTM), the matching process can be achieved. The work in [Majdik et al., 2013] uses ASIFT features to describe the images. The ASIFT algorithm aims to solve the problem of tilt matching caused by some descriptors, such as SIFT, ORB and SURF. Though the above descriptors perform well under scale scaling and rotation conditions, they can hardly detect the key points of oblique images and thus ASIFT is proposed to tackle this issue.

Local features are widely used in the task of image-based localization because of its variance properties under several conditions, such as illumination, viewpoint, cropping changes, etc. But in changing conditions, global features can have better performance than local features [Lowry et al., 2016].

#### 2.2.1.2/ GLOBAL FEATURES

Global features do not need a selection phase, but describe the image as a whole. These features usually need little computation time and can give the localization task good results. Table 2.3 gives a summary of the main global features and labels their reference. Though this is not a complete list, this can help the readers to have a basic knowledge of the global features.

Table 2.3: Summary of the main global features.

| Name | Reference |
|---|---|
| Principal components | [Gaspar et al., 2000] |
| Colour histograms | [Ulrich et al., 2000] |
| Gradient orientation histograms | [Kosecka et al., 2003] |
| Receptive field histograms | [Pronobis et al., 2006] |
| Gist | [Oliva et al., 2001] |
| BRIEF-gist | [Sünderhauf et al., 2011] |
| Fourier signatures | [Menegatti et al., 2004] |
| Normalized patches | [Milford et al., 2012] |
| Scanline intensity profile | [Milford et al., 2008] |
| Spherical harmonics | [Chapoulie et al., 2013] |
| WI-SURF | [Badino et al., 2012] |
| WI-SIFT | [Badino et al., 2012] |
| DIRD | [Lategahn et al., 2013] |
| WGOH | [Bradley et al., 2005] |
| OACH | [Junqiu Wang et al., 2006] |

Table 2.4: Summary of image-based localization approaches based on global features.

| References | Camera | Tasks | Environment | Feature |
|---|---|---|---|---|
| [Bradley et al., 2005] | Mono | Map + Loc | Outdoor | WGOH |
| [Sünderhauf et al., 2011] | Mono | SLAM | Outdoor | BRIEF-gist |
| [Arroyo et al., 2014b] | Omnidir | Map + Loc | Outdoor | Concatenation of LDB |
| [Arroyo et al., 2014a] | Stereo | Map + Loc | Outdoor | Concatenation of D-LDB |
| [Liu et al., 2012] | Mono | SLAM | Outdoor | Gist |
| [Prasser et al., 2006] | Omnidir | SLAM | Outdoor | Colour histograms |
| [Milford et al., 2008] | Mono | SLAM | Outdoor | Scanline intensity prof. |
| [Glover et al., 2010] | Mono | SLAM | Outdoor | Scanline intensity prof. |
| [Badino et al., 2012] | Mono | Map + Loc | Outdoor | WI-SURF |
| [Xu et al., 2014] | Mono | Map + Loc | Outdoor | WI-SURF |
| [Lategahn et al., 2013] | Mono | SLAM | Outdoor | DIRD |
| [Pepperell et al., 2014] | Mono | SLAM | Outdoor | Normalized patches |
| [Milford et al., 2012] | Mono | SLAM | Outdoor | Normalized patches |
| [Junqiu Wang et al., 2006] | Mono | Map + Loc | In + Out | OACH |
| [Ulrich et al., 2000] | Omnidir | Map + Loc | In + Out | Colour histograms |
| [Chapoulie et al., 2013] | Sphere | Map + Loc | In + Out | Spherical harmonics |

Table 2.4 lists some of the image-based localization approaches based on global features. As this thesis is concentrated in localization on outdoor environments only, the approaches in this table focus on outdoor conditions or can be applied in outdoor conditions. Based on the descriptors and the information processing methods, these approaches can be divided into three main categories [Garcia-Fidalgo et al., 2015]:

**(1) histograms.** The works of [Ulrich et al., 2000], [Bradley et al., 2005] and [Junqiu Wang et al., 2006] all use the compact histograms to describe an image for the image-based localization. The image in [Ulrich et al., 2000] is described by six 1D colour histograms, of which each three are extracted from RGB and HLS colour space, respectively. Weighted Gradient Orientation Histogram (WGOH) features are used in [Bradley et al., 2005]. In this work, images are segmented into different parts and on each part an 8-bin histogram of the gradient orientations is extracted. By concatenating all the resulted histograms, a WGOH feature is generated. Orientation Adjacency Coherence Histogram (OACH) in the work [Junqiu Wang et al., 2006] is proposed and based on the Histograms of oriented gradient (HOG) to deal with the coarse regions, such as edges and corners of an image.

**(2) Gist descriptor.** Gist [Oliva et al., 2001] is inspired by the human ability of recognizing images by a simple glance under certain environments and has been applied in several works recently. After generating the features by Gist, Principal Component Analysis (PCA) is applied to reduce the dimensionality [Liu et al., 2012]. As Gist can not well describe spherical images, [Chapoulie et al., 2013] proposes a novel descriptor based on spherical harmonics. By translating the local feature BRIEF to a global feature based on Gist, [Sünderhauf et al., 2011] proposed BRIEF-Gist feature. While bidirectional loops are ignored by BRIEF-Gist, the method of [Arroyo et al., 2014b] is proposed to solve this. Each panorama in this method is represented by concatenating different LDB (Local Difference Binary) binary descriptors of the segmented sub-panoramas. Then the LDB feature is advanced to D-LDB feature by adding disparity information in their work in [Arroyo et al., 2014a].

**(3) biologically-inspired approaches.** Inspired by the behaviours of living creatures, some approaches learn to process the information and solve the problem simulating these behaviours. One of the state-of-the-art work is RatSLAM [Milford et al., 2004]. Simulating the hippocampus structure of rats, RatSLAM builds continous attractor networks to

represent environment, which permits to a closing loop for mapping to correct odometry error. Its localization task is under indoor conditions work. Later, [Prasser et al., 2006], [Milford et al., 2008] and [Glover et al., 2010] propose advanced works based on RatSLAM, and solve the localization problem under outdoor conditions.

**(4) others.** In the other image-based localization approaches, [Badino et al., 2012] and [Xu et al., 2014] use the Whole Image SURF (WI-SURF) descriptor. This descriptor is extracted by applying the local descriptor SURF in the whole image based on the method of [Agrawal et al., 2008]. [Lategahn et al., 2013] uses an Illumination Robust Descriptor (DIRD) in their work, which is robust to the severe changes of illumination. SeqSLAM proposed in [Milford et al., 2012] uses normalized patches to extract sequence instead of a single image and achieves good performance under weather or season variations. SMART in [Pepperell et al., 2014] is proposed based on SeqSLAM, which changes the image matching method of SeqSLAM and uses the same descriptor, normalized patches.

The disadvantage of global features is their pose dependency. As neither of the global and local features are perfect, combining them is the choice in some works [McManus et al., 2014]. This is achieved by segmenting the images into regions and then using the global features to represent these regions.

### 2.2.2/ APPROACHES BY USING CNN-BASED FEATURES

Thanks to the advances of software and hardware, deep learning (DL) has known a development explosion and is successfully applied in computer vision and robotics. In the recent years, the advanced approaches of computer vision [Gu et al., 2018] and robotics[Sünderhauf et al., 2015a] have been used in localization to improve the performance. The model of Convolutional Neural Network (CNN) is most commonly applied in

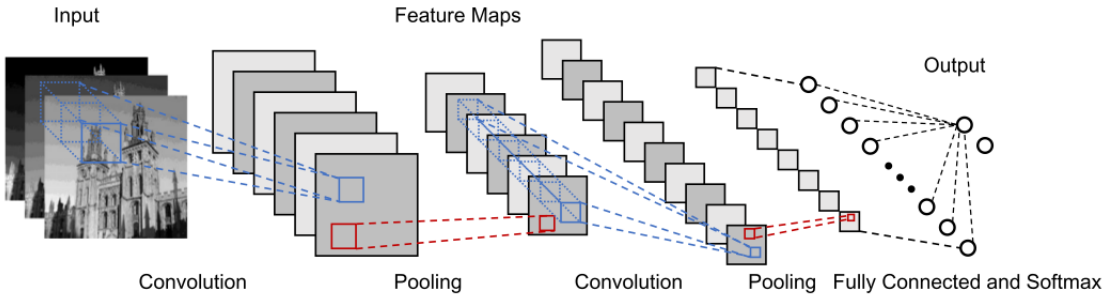



Figure 2.3: The typical architecture of CNN ([Zhang et al., 2021]).

the DL-based localization to generate the deep learned features of images. The typical architecture of a CNN is shown in Fig. 2.3, which includes convolutional layers, pooling layers, fully connected layers and a Softmax layer.

One of the famous early works of CNN model, LeNet, was proposed in 1998 [Lecun et al., 1998]. Due to the limitation of their complexity, CNNs have not drawn big attention until 2012, at which time AlexNet was proposed and won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). After then, CNNs make great progress and the model of VGG16, GoogleNet and ResNet have been proposed and widely applied into the localization tasks. The details of these CNN models will be given in Chapter 3.

In the early CNN-based localization works, features of image are generated directly by the pre-trained CNN models. Then considering the high dimensionality of the feature vectors, methods are proposed to process these CNN-based features before matching them. In addition to process the features, developing the CNN models, such as fine-tuning the pre-trained model or changing the CNN architecture, are other ways to improve the robustness of CNN-based features for the localization task.

Therefore, the CNN-based features can be divided into four categories: 1) off-the-shelf features from the pre-trained CNN model, 2) post-processed features from the pre-trained CNN model, 3) features from the fine-tuned CNN model, 4) features from new architecture CNN model.

**1) Off-the-shelf features from a pre-trained CNN model.** The early works using the features generated from a pre-trained CNN model can date back to 2014 [Chen et al., 2014], which uses the Overfeat model to get the features.

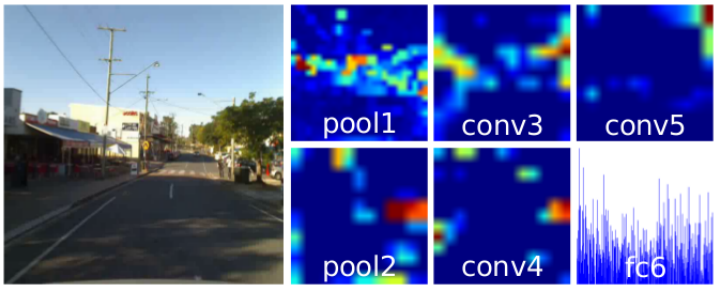



Figure 2.4: The features extracted from different layers of AlexNet network ([Sünderhauf et al., 2015a]).

Soon after, [Sünderhauf et al., 2015a] did a similar work with the pre-trained AlexNet model. This work is very inspiring as it finds that features obtained from different layers of a network can have robust performance under different conditions. As shown in Fig. 2.4, features of the same image can be extracted from different network layers and seen as holistic image descriptors. Though comparing the performance of these different features in the task of place recognition, a conclusion is made that features from middle layers are good at describing images with appearance changes, while features from the top layers have better show the effectiveness under viewpoint changing conditions.

Then the work of [Hou et al., 2015] proved that the CNN-based features are more robust than the hand-crafted features under severe lighting changing conditions through experiment comparisons. [BAI et al., 2018] used the features based on the pre-trained AlexNet to improve the global feature based localization work, SeqSLAM [Milford et al., 2012].

**2) Post-processed features from the pre-trained CNN model.** As the possibility of directly applying the features extracted from a pre-trained CNN model has been well exploited in the early localization and place recognition work, researches of post-processing the CNN features have been proposed. There is a simple improvement in the work of [Hou et al., 2015] by applying the $\ell_2$-normalization on the output CNN features. Then, the method in [Sharif Razavian et al., 2014] is proposed to reduce the dimensionality of the CNN features. As Principal component analysis (PCA) is the most commonly used method to reduce the dimensionality of vectors, [Sharif Razavian et al., 2014] combined the PCA with whitening (a method in deep learning to reduce redundancy of the input features) in the post-processing of CNN features. The work in [Zhang et al., 2017] is complex for taking the advantage of all the above methods, by combining not only $\ell_2$-normalization, but also the PCA compression and whitening method on the CNN features. These methods improved the performance of image-based localization step by step.

**3) Features from the fine-tuned CNN model.** Though the features based on the pre-trained CNN models have good performance in image-based localization, these models are not trained especially for the localization tasks. Fine-tuning these models on specific localization dataset can help in improving the robustness of CNN features for the image-based localization tasks. As an important step for image-based localization, place recognition method applied a fine-tuned CNN model firstly in [Lopez-Antequera et al., 2017] for extracting appearance-invariant features. [Chen et al., 2017a] proposed a dataset named

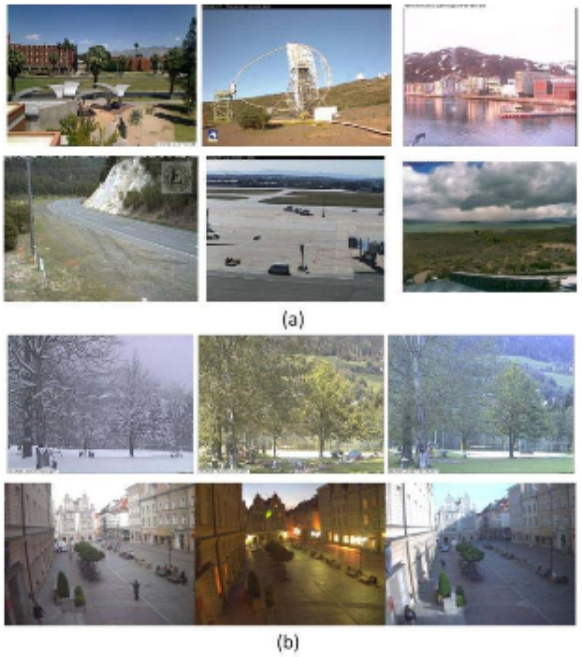



Figure 2.5: Example images of the dataset SPED ([Chen et al., 2017a]). (a) Images in this dataset have appearance diversity. (b) Images in the same row represent the same place with large condition changes at different times.

Specific PlaceEs Dataset (SPED) which has 2.5 million images under drastically conditions and viewpoint changes (as shown in Fig. 2.5). Then the pre-trained CaffeNet is initialized and fine-tuned on SPED, thus developing a new network, HybridNet. The features generated by HybridNet is discriminative under condition and viewpoint changing environments. [Radenović et al., 2019] proposed to select the training images automatically instead of doing it manually by the method of structure-from-motion (SFM).

**4) Features from new architecture CNN model.** Beside using the raw CNN architectures, new methods are proposed to design new CNN architectures to improve the image-based localization performance.

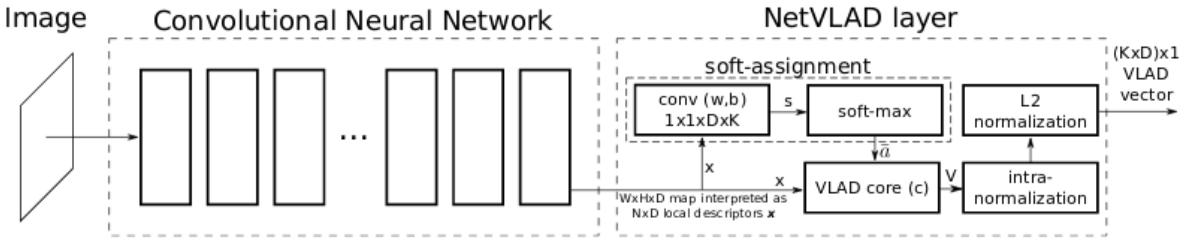



Figure 2.6: The architecture of NetVLAD [Arandjelovic et al., 2016].

Among these novel methods, a state-of-the-art architecture named as NetVLAD is pro-

posed in [Arandjelovic et al., 2016]. This is an end-to-end network based on a common CNN architecture. But at the end of the CNN framework, the authors add a NetVLAD layer which performs aggregation on the output CNN features instead of directly outputting them. The framework of NetVLAD can be seen in Fig. 2.6.

As shown in Fig. 2.7, an attention learning system used to evaluate the importance of each spatial location of the image is proposed in [Chen et al., 2018]. Through the feature maps $(M_1, M_2, M_3)$ of the input image generated from different layers of the pre-trained CNN model, their soft attention masks $(S_1, S_2, S_3)$ are calculated, respectively. Instead of fixed contextual regions focused by other methods, this attention system allows more flexibility in learning contextual regions to generate the masks. Then the final mask $G$ is calculated by fusing the masks obtained from multi layers. Depending on the mask $G$, the final features, corresponding to the importance of the spatial locations, can be generated.

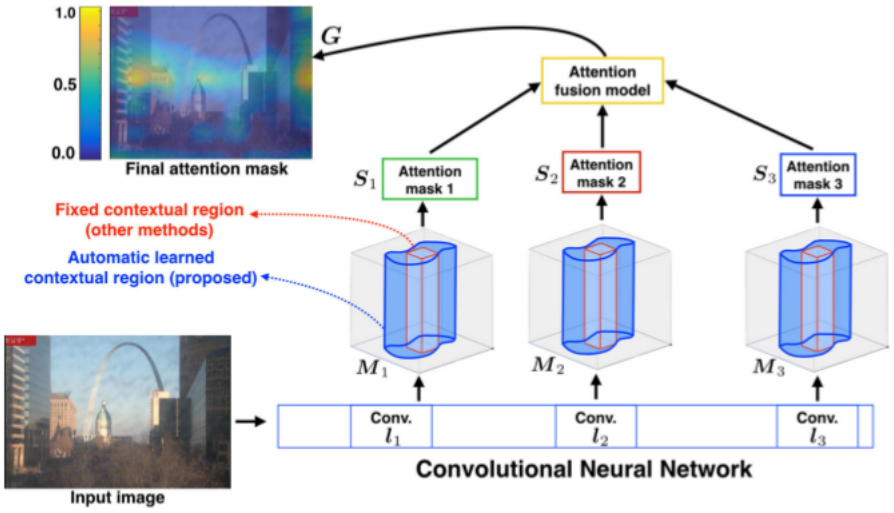



Figure 2.7: The attention learning system [Chen et al., 2018].

As Siamese and Tripet networks are common architectures to train the metric learning model, they are used in many image-based localization researches [Hausler et al., 2019] [Xin et al., 2019] [Zhao et al., 2019] in which the localization is seen as a distance metric leaning process.

Other similar methods are listed in Table 2.5 as well as the above generating CNN-features based methods.

Table 2.5: Summary of the CNN-based features.

| Reference | Feature |
|---|---|
| [Chen et al., 2014] | Off-the-shelf features (pre-trained CNN model) |
| [Sünderhauf et al., 2015a] | Off-the-shelf features (pre-trained CNN model) |
| [Hou et al., 2015] | Off-the-shelf features (pre-trained CNN model) |
| [BAI et al., 2018] | Off-the-shelf features (pre-trained CNN model) |
| [Hou et al., 2015] | Post-processed features (pre-trained CNN model) |
| [Sharif Razavian et al., 2014] | Post-processed features (pre-trained CNN model) |
| [Zhang et al., 2017] | Post-processed features (pre-trained CNN model) |
| [Lopez-Antequera et al., 2017] | fine-tuned CNN features |
| [Chen et al., 2017a] | fine-tuned CNN features |
| [Radenović et al., 2019] | fine-tuned CNN features |
| [Arandjelovic et al., 2016] | new architecture CNN features |
| [Chen et al., 2018] | new architecture CNN features |
| [Hausler et al., 2019] | new architecture CNN features |
| [Xin et al., 2019] | new architecture CNN features |
| [Zhao et al., 2019] | new architecture CNN features |
| [Kim et al., 2017] | new architecture CNN features |
| [Merrill et al., 2018] | new architecture CNN features |
| [Chen et al., 2015] | new architecture CNN features |

### 2.2.3/ APPROACHES BASED ON SYNTHETIC FEATURES

Even making use of the most powerful CNNs, recognizing places under strong appearance changes, for example, due to change of seasons, viewpoints, illumination and background modifications such as buildings, is still a challenging problem without satisfactory performance. This is due to a lack of appropriate training data. Training a model for a specific place recognition task, such as place recognition across strong changing appearance, needs hundreds of thousands of appropriate images obtained from different positions across changing conditions, to enable the robustness to illumination, viewpoint, appearance varieties. However, it is very difficult and expensive to gather this kind of data. Different from most approaches so far focusing on improving the feature representations, these works [Torii et al., 2015] [Porav et al., 2018] [Anoosheh et al., 2019] aim at modifying the input images and generating appropriate synthetic images.

The observation in [Torii et al., 2015] demonstrated that when the query and reference images are in the same viewpoint, the results of severe appearance changing place recognition can be better. Therefore, the place recognition approach in this paper which is named DenseVLAD generates synthetic novel views in an off-line stage before computing the densely extracted descriptors of query and reference images for matching and

retrieving. Based on the geo-tagged images of the dataset, densevlad generates a set of additional virtual viewpoints of these images. Examples of generated images can be seen in Fig. 2.8.

This view synthesis method mainly includes two steps. Firstly, candidate camera positions are selected from a regular $5m \times 5m$ grid on the map that covers the original camera position. To avoid strong artifacts in the generated views, virtual camera positions must be within 20 meters distance from the original camera position. Then the synthetic views at the virtual camera positions are generated based on the street-view panoramas and their corresponding piece-wise planar depth maps downloaded from Google maps[1]. For each query image and the synthesized views with the same viewpoint as the query image, the descriptors are extracted and then matched by finding the nearest descriptors across same viewpoints. This strategy shows an improvement compared to the results from the same matching procedure for the query image and synthesized view across different viewpoints.


Figure 2.8: Generating the synthesized image with the same viewpoint as the query image based on geo-tagged Google map. [Torii et al., 2015]. The dots and arrows in (d) indicate the camera positions and view directions.

With the development of GANs, robust GAN-based image-to-image translation methods [Anoosheh et al., 2018] [Choi et al., 2018] [Ignatov et al., 2018] [Liu et al., 2017] [Zhu et al., 2017] make it possible to generate drastically appearance changing synthesized images, e.g conditions transform day to night, summer to winter etc.

[1] http://maps.google.com/help/maps/streetview/.

[Porav et al., 2018] is the first work that applies synthetic images from nighttime images to daytime images for night visual place recognition. Based on previous work by [Linegar et al., 2015] [Churchill et al., 2013] [Zhu et al., 2017], this paper proposed a feature detector and descriptor pipeline in the process of the two-stages image translation network training.

In the first stage, two generators are built based on UResNet [Guerrero et al., 2018], whose architecture is a combination of UNet [Ronneberger et al., 2015a] and residual (ResNet) [He et al., 2016] modules. $G_{AB}$ transforms the original image (condition A, such as summer) into a synthetic image (condition B, such as winter), and L1 loss is applied between these two kinds of images. $G_{BA}$ transforms the synthetic image (condition B) into the reconstructed image (condition A). L1 loss is applied between the SURF feature maps of the input image and the reconstructed image, and between the dense per-pixel SURF feature maps of these two kinds of images, separately.

In the second stage, a dataset of aligned day and night images is used to train the $G_{AB}$ and $G_{BA}$ separately which makes it possible for the generators to learn certain feature transformations that might be ignored in the first stage. Two kinds of feature maps, SURF


Figure 2.9: Example images of the image translating results [Porav et al., 2018].

feature maps and the dense SURF feature maps, are generated for both the aligned target images and the generated ones. The L1 loss is used between the SURF feature maps of the two kinds of images firstly, and then is used between their dense SURF feature maps.

For dataset, this work ([Porav et al., 2018]) selected 6 traversals from the Oxford RobotCar dataset which contains 5 condition pairs: day-night, day-snow, day-dawn, day-sun and day-rain. By transforming the non-day weather condition from each condition into day (examples can be seen in Fig. 2.9), the place recognition results using the original day and synthesized day images have get an improvement compared to the results using both original images.

The work in [Anoosheh et al., 2019] also proposed a synthesized image based day and night place recognition method, known as ToDayGAN. Firstly, it trains an image translation model targeting in day and night image domains and translates the input query images from night to day. Then the features of both translated images and reference images are generated by the method of DenseVLAD. Through nearest neighboring search, a closest matching reference image of each query image is obtained, and the day-night localization problem is solved.

The image-translation model in ToDayGAN is based on the image-translation model of ComboGAN [Anoosheh et al., 2018], yet modifies the discriminators of ComboGAN to fit the night-to-day problem. Following WESPE [Ignatov et al., 2018], ToDayGAN uses three discriminators per domain - one for texture, one for color and one for gradients -

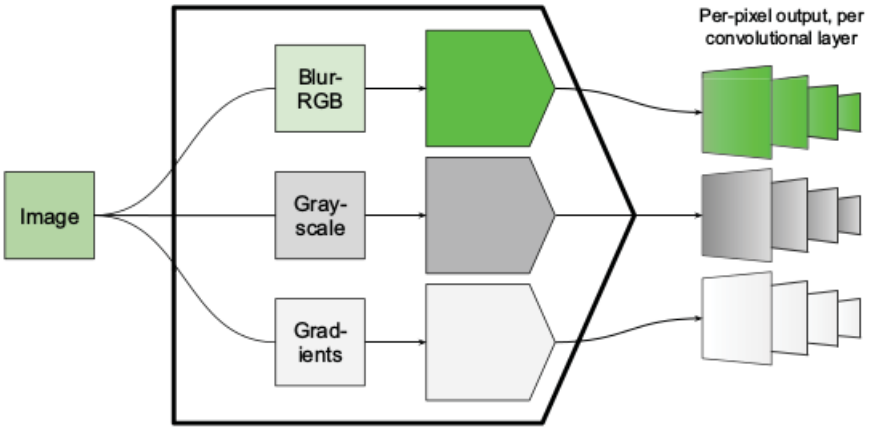



Figure 2.10: The framework of the three discriminators per domain used by ToDayGAN [Anoosheh et al., 2019].

rather than just one, as shown in Fig. 2.10. The architecture and hyperparameters of each discriminator are identical and the losses are averaged equally. The discriminator focusing on texture takes the luminance of the input image. The discriminator for color is as in WESPE, taking the RGB image after the application of a 5×5 Gaussian kernel. The last discriminator focusing on gradients of the input image is a novel contribution in ToDayGAN. It downsamples the input image by a 1×1 stride-2 convolution and convolves it with a $[-1\ 0\ 1]$ kernel for $x$-direction gradients and its transpose for the $y$ gradients.

The images of Oxford RobotCar dataset are used for training and evaluating the model ToDayGAN. A subset of 6954 day images and 6666 nighttime images of Oxford RobotCar dataset are selected for training and two query sets -one set of 438 night images, and a second set of 440 images captured at night during rain - are selected for testing.


Figure 2.11: Examples of the night to day translation by ToDayGAN [Anoosheh et al., 2019].


Figure 2.12: Examples of the transformed images from low-quality night images by ToDayGAN.

From high quality nighttime images, the translated daytime images by ToDayGAN have

a good quality, as shown in Fig. 2.11. But facing to low-quality nighttime images, even from the same dataset and similar environments with the query image used in ToDayGAN itself, ToDayGAN cannot achieve good translation performance, as illustrated in Fig. 2.12.

## 2.3/ APPROACHES FOR FEATURE MATCHING STAGE

After getting the robust features by a feature representation method, it is necessary to match these features to find the reference image that best corresponds with the query image. Feature matching is in fact a feature similarity searching process, which can be divided into two parts based on the similarity measure [Zhang et al., 2021]: (1) global similarity, (2) region- or landmark-based similarity.

### 2.3.1/ GLOBAL SIMILARITY BASED MATCHING

As discussed in Chapter 2.2, when the features are extracted from the whole image, whether it is represented by the hand-crafted or by CNN-based features, the features will be described as a single vector. In this way, the distance computation between the feature vectors is often used to measure the features similarity. As a very simple metric, Euclidean distance [Chen et al., 2014] [Hou et al., 2015] [Arandjelovic et al., 2016] is used in many works, as also in our work in this thesis. Then, a normalized Euclidean distance is introduced in [Lopez-Antequera et al., 2017]. [Zhang et al., 2017] refines the Euclidean distance to limit its value in [0,1].

Beside the Euclidean distance, cosine distance is another common metric used for matching. As the work of [Lopez-Antequera et al., 2017], a normalized version of cosine distance is proposed in [Naseer et al., 2018] to get the confusion matrix.

However, for the large scale or high-dimensional feature matching, the above brute-force methods cannot satisfy the computation need, thus more similarity search methods are proposed to handle this case. Indeed, some pre-processing methods are used to reduce the dimensionality of feature vectors. Locality sensitive hash (LSH) is used in [Sünderhauf et al., 2015a] to improve the rapidity without efficiency loss. Hashing methods are also used to improve the matching speed for the localization problem. The deep supervised hashing method introduced in [Wu et al., 2019] can achieve real-time place

matching under drastically lighting or viewpoint changing conditions.

### 2.3.2/ MATCHING BY METHOD OF REGION OR LANDMARK-BASED SIMILARITY

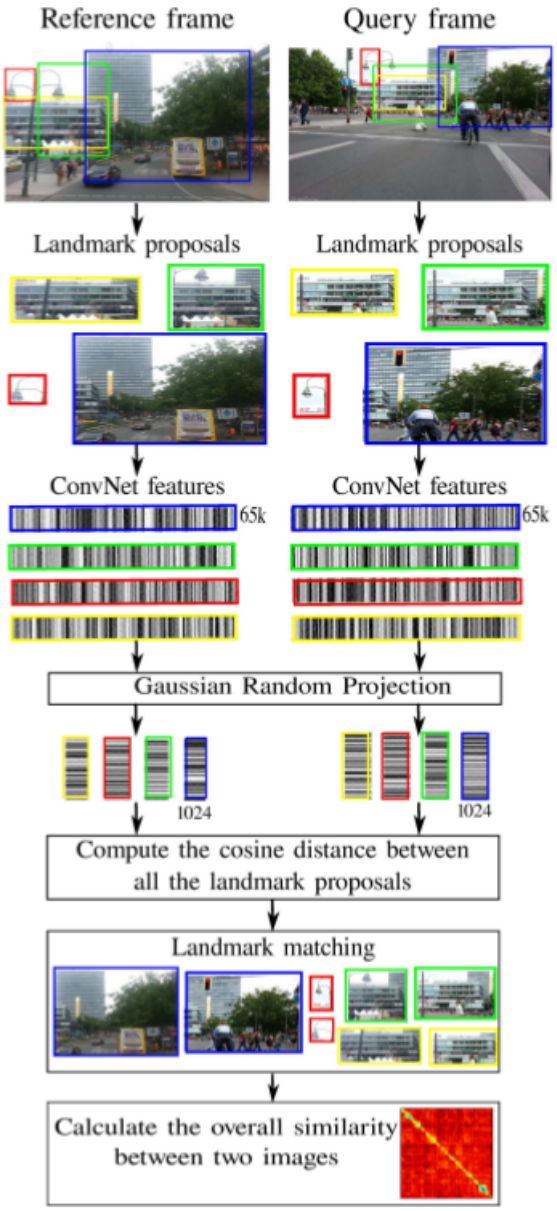



Figure 2.13: The framework of the landmark-based place matching [Sünderhauf et al., 2015b].

For the local features from salient regions or landmarks, the matching process is much more complex than for the global features. It does not only need to compute the similarity of the matched landmarks but also the overall similarity of the query and reference images. The work of [Sünderhauf et al., 2015b] illustrates the complete framework of

landmark matching. As shown in Fig. 2.13, the dimension reducing method - Gaussian Random Projection - is firstly applied on the CNN-based landmark features. Then the compressed (1024 dimensions) landmark features are matched by nearest neighbouring searching using cosine distance as the metric. If two landmarks, the nearest neighbour of one landmark is the other, and vice versa, they will be seen as the matched landmark pair. Then in the landmark matching step, the shapes of the bounding boxes of this landmark pair are scored and matched. Finally, the overall similarity between two images is calculated and the only best matching image is found according to the similarity score of the query and reference images.

A similar work can be found in [Xin et al., 2019] and [Chen et al., 2017b]. [Xin et al., 2019] uses the weight similarity instead of the shape similarity for the landmark matching. And the work of [Chen et al., 2017b] trains a vocabulary to represent the landmark features. Cosine distance is used in both two works as the matching metric.

## 2.4/ CONCLUSION

In this chapter, a survey of recent works on visual based vehicle localization as well as place recognition which can be seen as a core step of visual based localization is introduced.

As the outdoor environment is very complex, one of the difficulties to which visual based vehicle localization faces is the severe appearance changes of the scenes under the changing of season, weather, illumination, viewpoint, traffic conditions, etc. Therefore, finding robust features to represent the images and effective matching methods to retrieve the correct image from the reference dataset are two problems that determine the precision of the localization task. Many of the localization approaches proposed are based on feature representation and feature matching.

Firstly, the image representation based vehicle localization researches are reviewed in this chapter. Hand-crafted features are useful descriptors to represent images in the early work. There are two ways to generate the hand-crafted features, one is to select the salient regions of a image and describe them by local features, the other is to describe the image as a whole using global features. With the development of the deep learning technology, the CNN-based features show their robustness for visual based localization

problem. The localization approaches based on CNN features vary from the off-the shelf features, post-processed features, fine-tuned CNN features to the CNN features using CNN models which have new architectures.

Then, the matching based visual localization approaches are also introduced in this chapter. The nearest neighbouring searching methods using the Euclidean distance or cosine distance are the most commonly matching methods for global features. Then the feature compressed methods and the hashing methods are often used in the feature matching to improve the time efficiency. For the region or landmark-based features, the matching process is more complex.

Introduced in the next four chapters, the contribution works in this thesis deal with feature representation based localization approaches which aim to generate robust features for recognizing a place.

# II

# CONTRIBUTION

# 3

# Overview of the Proposed Visual Place Recognition Framework

This thesis focuses on visual place recognition, which is a core process for visual-based localization tasks, as shown in 1.2 . The proposed approaches (chapters 4, 5 and 6) in this thesis are developed under the same place recognition framework, which is introduced in this chapter. For solving the place recognition problem, a feature representation method should be applied first to get distinguishable features of each reference and query images. Then an image retrieval process is needed to select the appropriate reference image which represents the same place as the query image. Therefore, the place recognition system including both the featurization process and the image retrieval process will be explained in details.

Beside the place recognition process, the common experimental setup that will be applied in chapters 4, 5 and 6 to evaluate the proposed approaches will be described in this chapter. It includes the evaluation datasets, data pre-precessing and evaluation metrics and baseline. The query and reference images are obtained from three open source datasets: St. Lucia[1], CMU[2] seasons and Oxford RobotCar[3]. Instead of directly using the selected images, an image pre-processing step cannot be avoided. After getting the best matching reference image of the query image by the place recognition system, metrics are needed to judge if the matching is correct or not. Baseline approaches are used to evaluate and compare the performance of our proposed approaches.

[1] http://asrl.utias.utoronto.ca/~mdw/uqstluciadataset.html
[2] https://www.visuallocalization.net/datasets/
[3] https://robotcar-dataset.robots.ox.ac.uk/datasets/

## 3.1/ IMAGE RETRIEVAL BASED PACE RECOGNITION METHOD

In this section, the framework of the image retrieval based place recognition is detailed firstly. Then the CNN models used for extracting the CNN-based features of an image are explained.

### 3.1.1/ THE FRAMEWORK OF THE PACE RECOGNITION METHOD

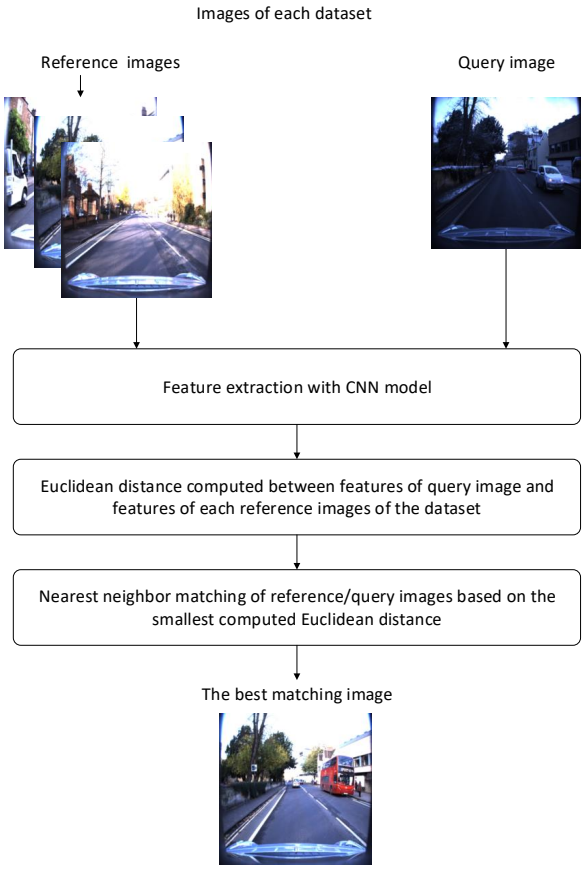



Figure 3.1: The process of the place recognition system.

The place recognition method (as seen in Fig. 3.1) is based on image retrieval. When a robotcar revisits a place and takes its appearance image by the sensor mounted on this car, this image representing the place can be seen as a query image. For the image retrieval based place recognition, there must be a priori map which contains the appearance images of the visited places by the robotcar. The images contained in the map are seen as the reference images, as shown in Fig. 3.1. The objective of image retrieval

based place recognition is to select from the reference images the best matching image that has the most similar feature as the query image. As key step of localization, once the place recognition is completed, the location of the current place can be learned by reading the GPS data of the best matching reference image.

Among the different kinds of features generated for localization tasks as described in section 2.2 of Chapter 2, CNN-based features are used in this thesis. The CNN-based features of both query images and the reference daytime images are generated by a featurization process. The featurization process is very simple as to get the deep features of the input image by the well-trained CNN models which are good at extracting image features under traffic conditions. The details of the selected models will be presented in the following.

As seen in Fig. 3.1, after getting the feature representation of each image, it needs to find some methods to match these features. As the CNN-based features are global features, one of the most simple methods, Euclidean distance of the two feature vectors is used as the metric to compute the similarity of each pair of two images. When the similarity scores between the query image and each of the reference image are calculated, nearest neighbour searching is applied to find the only best matching reference image according to the highest similarity score. The place recognition process is well completed by finding the best matching reference image.

### 3.1.2/ THE CNN MODELS USED FOR EXTRACTING FEATURES

To prove the effectiveness of our proposed methods, which are not relying on specific CNN features, four different pre-trained CNN models (VGG19[4], NIN_-ImageNet[5], AlexNet[6], bvlc_GoogLeNet[7]) are used in this paper. According to [Sünderhauf et al., 2015a], features from different layers of the CNNs have different robustness against appearance changes. Therefore, when the four network models are used in the proposed approaches, it is always that the layer conv5_1 in VGG19, conv3 in NIN_ImageNet, conv4 in AlexNet and pool3/3×3_s2 in bvlc_GoogLeNet are used to generate the CNN features of images.

[4] https://github.com/davidgengenbach/vgg-caffe
[5] https://github.com/dsys/nin-imagenet/tree/master/model
[6] https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet
[7] https://github.com/BVLC/caffe/tree/master/models/bvlc_googlenet

The four CNN models are detailed as below:

- AlexNet. AlexNet is a large and deep CNN proposed in [Krizhevsky et al., 2012]. It was trained to classify the images in the ImageNet LSVRC-2010 contest. It was then improved and achieved the champion's in the LSVRC-2012 competition.

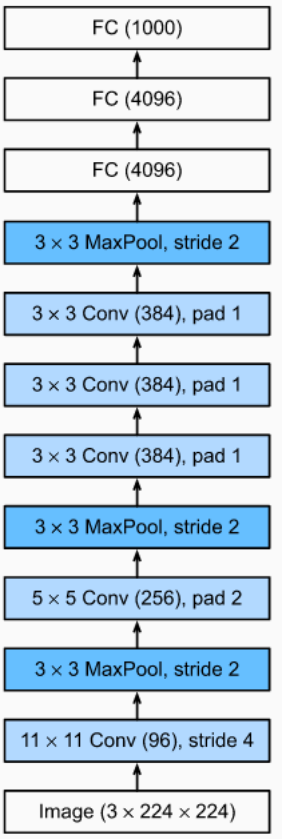



Figure 3.2: The architecture of AlexNet [Zhang et al., 2020].

As a pioneer, AlexNet has set the tone for subsequent CNN and even R-CNN networks. The architecture of the network contains five convolutional layers and three fully-connected layers, as shown in Fig. 3.2. But this is not the contribution of AlexNet; however it implemented new ideas to CNNs:

1) AlexNet uses Rectified Linear Units (ReLU) instead of the tanh function.

2) AlexNet uses LRN(Local Response Normalization). LRN normalizes the adjacent areas of the data without changing the size and dimensions of the data. Through achieving local suppression, the generalization ability of the model has been improved.

3) AlexNet applies Overlapping Pooling. The conventional CNN summarizes the

neighborhoods by adjacent pooling units with no overlapping. However, by utilizing overlap, error rates on the test data can have a small reduce. And models with overlapping are a litter harder to overfit during training process.

4) AlexNet introduces "dropout" [Hinton et al., 2012] in the first two fully-connected layers: fc6 and fc7. With dropout, the models are difficult to overfit and have stronger generalization ability. It can also decrease the complexity of the networks and increase the speed of computation.

5) AlexNet proposes some data augmentation schemes. The input of AlexNet is $224 \times 224$ image, but the pre-processed images have the size of $256 \times 256$, this is because a $224 \times 224$ sliding window is used to extract random patches from the $256 \times 256$ images. This amplifies the training set by 2048 times.

- VGG19. VGG [Simonyan et al., 2014] is proposed by Visual Geometry Group and is the improved version of their related work in the ILSVRC-2014 competition. VGG has two architectures, VGG16 and VGG19, as shown in Fig. 3.3 (D) and (E). The difference between VGG16 and VGG19 is the depth of the network. The pre-trained VGG19 model on ImageNet dataset is chosen in this thesis and can be easily gotten from the open source online[4].

  Compared to AlexNet, VGG19 uses a stack of $3 \times 3$ convolutional (conv) layers with stride 1 throughout the whole net instead of using relatively large conv layers, e.g. $11 \times 11$ or $7 \times 7$. It can be seen that using two $3 \times 3$ convolutional layers can achieve an effective receptive field of $5 \times 5$, the same result as using one $5 \times 5$ convolutional layer; using three $3 \times 3$ convolutional layers can achieve an effective receptive field of $7 \times 7$, the same result as using one $7 \times 7$ convolutional layer. In this way, VGG19 improves the depth of the net, while guaranteeing the same effective receptive field as the previous CNNs. It is proved that increasing the depth of the conventional CNN architecture can achieve the state-of-the art performance.

- NIN_ImageNet. NIN_ImageNet is also chosen to extract deep features of input images. It is a 4 layer NIN (Network in Network) [Lin et al., 2013] model trained on Imagenet dataset and its framework is shown in Fig. 3.4. Compared to conventional CNNs such as AlexNet, the NIN model changes its deep network structure to be smaller yet faster to train and has slightly better performance.

| ConvNet Configuration | | | | | |
|---|---|---|---|---|---|
| A | A-LRN | B | C | D | E |
| 11 weight layers | 11 weight layers | 13 weight layers | 16 weight layers | 16 weight layers | 19 weight layers |
| input ($224 \times 224$ RGB image) | | | | | |
| conv3-64 | conv3-64 **LRN** | conv3-64 **conv3-64** | conv3-64 conv3-64 | conv3-64 conv3-64 | conv3-64 conv3-64 |
| maxpool | | | | | |
| conv3-128 | conv3-128 | conv3-128 **conv3-128** | conv3-128 conv3-128 | conv3-128 conv3-128 | conv3-128 conv3-128 |
| maxpool | | | | | |
| conv3-256 conv3-256 | conv3-256 conv3-256 | conv3-256 conv3-256 | conv3-256 conv3-256 **conv1-256** | conv3-256 conv3-256 **conv3-256** | conv3-256 conv3-256 conv3-256 **conv3-256** |
| maxpool | | | | | |
| conv3-512 conv3-512 | conv3-512 conv3-512 | conv3-512 conv3-512 | conv3-512 conv3-512 **conv1-512** | conv3-512 conv3-512 **conv3-512** | conv3-512 conv3-512 conv3-512 **conv3-512** |
| maxpool | | | | | |
| conv3-512 conv3-512 | conv3-512 conv3-512 | conv3-512 conv3-512 | conv3-512 conv3-512 **conv1-512** | conv3-512 conv3-512 **conv3-512** | conv3-512 conv3-512 conv3-512 **conv3-512** |
| maxpool | | | | | |
| FC-4096 | | | | | |
| FC-4096 | | | | | |
| FC-1000 | | | | | |
| soft-max | | | | | |

Figure 3.3: ConvNet configurations [Simonyan et al., 2014]: Columns D and E stand for VGG16 and VGG19, respectively.

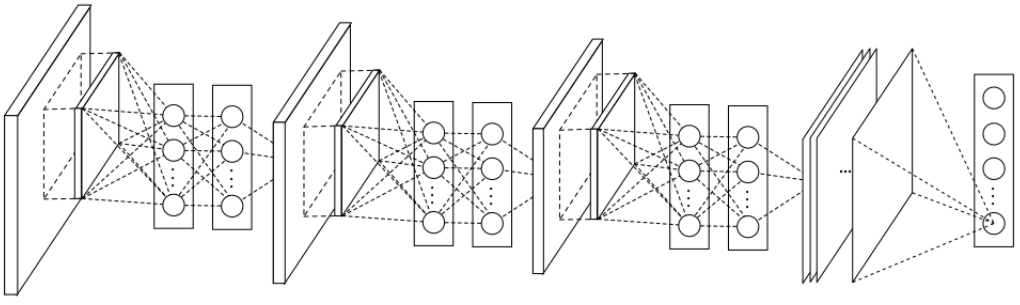


Figure 3.4: The framework of Network In Network [Lin et al., 2013]. NINs include three mlpconv layers and one global average pooling layer.

Classic CNNs are composed of convolutional layers, pooling layers and fully connected layers. The convolutional layer consists of linear convolutional filters and nonlinear activation functions to generate the feature maps. But this linear convolutional is not sufficient for getting the highly non-linear latent concepts which are needed for achieving good representation of the input data. Therefore, an overcomplete set of filters is used in the conventional CNN to cover different kinds of latent concepts. However, having too many filters results in too many parameters and too large network. As in CNN, a higher level concept is generated by combining the lower level concepts from the previous layers in different ways. Therefore, the authors of the NIN model argue that it is better to do the complex computation on each local patch instead of the higher level concepts and propose an advanced convolution layer: MLP convolutional (mlpconv) layer.

Besides, the fully connected layers are always used as the last layers of a conventional convolutional neural network and are prone to overfitting, thus damaging the ability of the network. To tackle this, NIN model uses the global average pooling instead of the fully connected layers.

Applying these two innovations, NIN model generates the feature map of input data in the last mlpconv layer. Then without fully connected layers, the resulting vector is gotten by taking the average of each feature map and then fed into the softmax layer. As benefit of this, NIN boosts the performance of the network as well as reducing the training time.

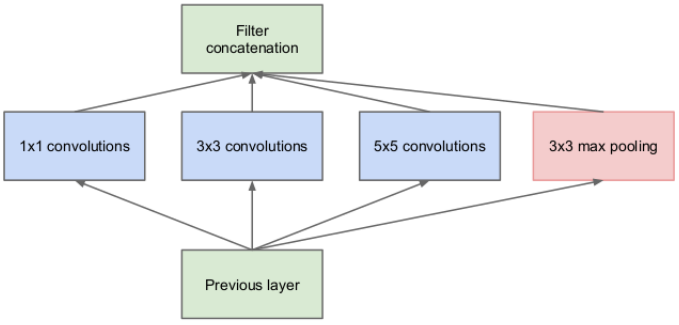
- bvlc_GoogLeNet. This model was trained by Sergio Guadarrama[8] and it is a replication of the work in [Szegedy et al., 2015].

  In order to break network symmetry and improve learning ability, traditional CNNs used random sparse connections. However, the computational efficiency of computer software and hardware for non-uniform sparse data is very poor, so the fully connected layer is re-enabled in AlexNet in order to better optimize parallel operations.

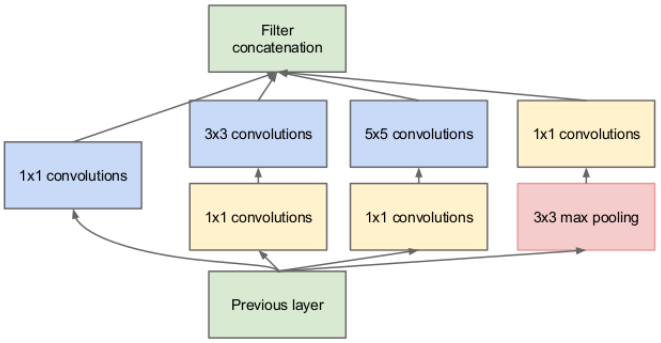
  Therefore, finding a new method that can not only maintain the sparsity of the network structure, but also use the high computational performance of the dense matrix is a direction that needs to be explored. A large amount of literature shows that

[8] https://github.com/BVLC/caffe/tree/master/models/bvlc_googlenet.

sparse matrices can be clustered into denser submatrices to improve computational performance. Thus GoogLeNet proposes the Inception architecture (as seen in Fig. 3.5) which takes advantage of dense components to approximate the optimal local sparse structure.




(a) Inception module, naïve version




(b) Inception module with dimensionality reduction

Figure 3.5: Inception architecture [Szegedy et al., 2015].

Inspired by NIN, the naive version of Inception uses a large number of 1x1 convolutional layers to improve the expressive ability of the network, while reducing the output dimension effectively. And the innovation of Inception is that it uses multiple convolutional layers of different sizes at the same time to capture information of different sizes in a structured way, which greatly reduces the amount of parameters and calculations.

Based on the problems that limit the performance of deep neural networks, continuous improvements of Inception architecture have been extended from the naive version to the fourth version: 1) Large parameter space, easy to overfit, and limited training data set; 2) Complex network structure and insufficient computing resources make it difficult to apply; 3) The deep network structure is prone to gradient dispersion, and decreases the performance of the CNN model.

## 3.2/ Overall Experiments Setup

In this part, the overall experiments setup which includes the evaluation dataset, the pre-processing of the selected data, and the evaluation metric and baseline approaches for comparison will be explained one by one.

### 3.2.1/ Evaluation dataset

Images which cover different conditions used to evaluate the proposed approaches in this thesis are selected from three datasets:

- St Lucia Dataset: The St Lucia Dataset [Warren et al., 2010] was captured in a suburb around the University of Queensland's St Lucia campus. The car passes through some routes several times in the same day and take pictures of the road-side scenery. For these continuous repeated routes, 227 images are selected to represent the places the car passes at the same time (as shown in the following part). Because of being taken in the same day, the images of the same place do not have a distinct variation in the appearance, in contrast to the sequences selected from the Oxford RobotCar dataset. An example of images of some places and the trajectory of this dataset can be seen in Fig. 3.6.

(a) Example of places in this trajectory

(b) Trajectory of St Lucia dataset

(c) The same places in this trajectory but taken in different times of the day

Figure 3.6: Vehicle path for the St Lucia dataset and the appearance of some example places at different times.

- CMU Seasons Dataset: The CMU Seasons dataset is a part of CMU Visual localization dataset created by *Badino et al* [Hernan Badino et al., 2011]. The images of 17 locations are captured under different conditions. Considering the appearance variations, only one location is selected and 207 images captured in this location are used in this thesis, as shown in Table 3.1. The images in CMU are obtained by both left and right mono cameras installed on a car. The sequence used in this thesis is taken from the left camera. The vehicle path of this dataset is not given, some example images captured at different times in the same place are shown in Fig. 3.7.

(a) Example places in CMU dataset

(b) The same place captured at different times

Figure 3.7: The appearance of some example places at different times.

- Oxford RobotCar Dataset: The Oxford RobotCar Dataset [Maddern et al., 2017] is composed of over 100 different sequences of Oxford. In over a year time, the Robotcar traveled through fixed routes in changing weather, seasonal and traffic conditions, even including constructions and roadworks. Among them, sequences captured at two trajectories under different weather and illumination are selected in this thesis, which will be described in the following part. One of the trajectory in this dataset and some of the places represented by images obtained at different times can be seen in Fig. 3.8. Those images chosen were captured by Bumblebee XB3 center sensor and were in unrectified 8-bit raw Bayer format. Before putting them into CNN networks, they were firstly converted to RGB.

(a) Example places in this trajectory

(b) One of the trajectory (2014/05/06) of Oxford dataset

(c) The same places in this trajectory at different times

Figure 3.8: One of the vehicle paths of Oxford dataset and the appearance of some example places at different times.

In fact, these datasets provide the 3DoF or 6DoF pose information instead of the GPS information, so the place of the query image can not be located directly by reading the geo-label of its best matching reference image. But these datasets provide different images under variable conditions that it can satisfy the dramatically appearance changes of the same places. Therefore, it is very suitable for place recognition experiments to prove the effectiveness of the proposed methods under even severe conditions. Once their effectiveness has been proved, applying the methods to the geo-tagged datasets, will not damage their place recognition performance, and the location can be easily achieved through the pose of the correct matching image.

For the place recognition experiments, different sequences are selected from the three datasets which deal with mild, moderate and severe variations in environmental conditions, containing illumination, traffic and pedestrians, weather and seasonal changes. Each sequence is a subset of two traversals of one fixed route which were taken over different times of a day or different seasons. Images from one traversal are used as query images, and the images of the other traversal are used as the reference images.

For evaluating the dynamic object removal place recognition method which will be introduced in Chapter 4, five daytime sequences are chosen as the experimental dataset.

The sequence day1 from the Oxford RobotCar dataset has little variations in traffic and weather, and its query and reference images were captured both in summer which means having no season changing. 258 images, including query and reference images, are

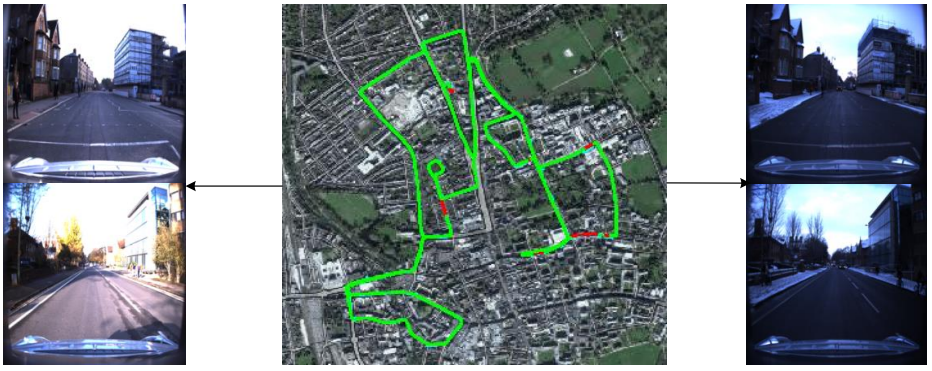
Table 3.1: The description of evaluation with daytime sequences.

| Sequences | Datasets | No. of frames | Variations in traffic | weather | season | Resolution Original | Resized | Recorded |
|---|---|---|---|---|---|---|---|---|
| day1 | Oxford RobotCar | 258 | Mild | Mild | None | 1280×960 | 224×224 | 2014/05/06 2014/06/23 |
| day2 | CMU | 207 | Severe | Moderate | Moderate | 1280×960 | 224×224 | non label |
| day3 | Oxford RobotCar | 200 | Severe | Severe | None | 1280×960 | 224×224 | 2014/11/18 2015/02/03 |
| day4 | St.Lucia | 227 | Mild | None | None | 640×480 | 224×224 | non label |
| day5 | Oxford RobotCar | 282 | Severe | Moderate | Severe | 1280×960 | 224×224 | 2014/07/14 2015/02/03 |

selected from some of the routes in the trajectory of Fig. 3.8 (b), which represents about 900 meters. And some of the example query images are shown in Fig. 3.8 (a), while Fig. 3.8 (c) shows the example reference images captured in the sequence 2014/06/03 of the Oxford RobotCar dataset.

For the CMU sequence, named day2, its two traversals were occurred in summer and fall with moderate changes in illumination and seasons, but with severe changes in the presence of dynamic objects, as can be seen in Fig. 3.7 (a) and (c) which show the query and reference images, respectively.

The sequence day3 is also selected from the Oxford Robotcar dataset. There are 200 query and reference images in this sequence which is captured in winter conditions and has severe appearance variations in weather and dynamic objects, as shown in Fig. 3.9.



(a) Examples of the query images.

(b) The trajectory of sequences 2014/11/28 and 2015/02/03 of the Oxford RobotCar dataset. The vehicle path of day3 is part of this trajectory.

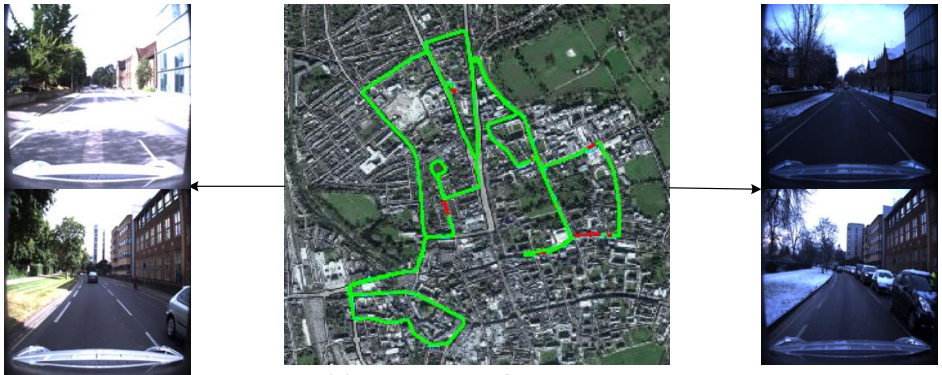
(c) Examples of the reference images.

Figure 3.9: The trajectory of sequence day3 and some examples of the query and reference images which represent the same places.

Besides the three sequences, other two sequences from St. Lucia and Oxford RobotCar datasets are also used to prove the effectiveness of the proposed place recognition

methods in Chapters 4 and 5.

The query and reference images from St. Lucia (sequence day4) are taken at different times in the same day, thus having similar appearance and few changing, as seen in Fig. 3.6 (a) and (c). While the two traversals of the sequence day5 are taken in summer and winter, thus having severe traffic, season changes and moderate changes in weather, as seen in Fig. 3.10. There are 282 images, including query and reference images, in day5 which are captured between two points of the path separated by a distance of about 1 km. Though the reference images of day5 and day3 are selected from the same sequence, different routes are chosen in the two sequences.



(a) Examples of the query images.

(b) The trajectory of sequences 2014/07/14 and 2015/02/03 of the Oxford RobotCar dataset. The vehicle path of day5 is part of this trajectory.

(c) Examples of the reference images.

Figure 3.10: The trajectory of sequence day5 and some examples of the query and reference images which represent the same places.

The place recognition approach proposed in chapter 6 tackles the day and night place recognition problem. The key point is to see the performance of our approach in night conditions, where images obtained from three different night conditions in the same route are used as query images and compared with the same daytime images in this route, respectively. Therefore, four sequences from RobotCar dataset are used, including one reference traversal in overcast day condition and three query traversals run at night, night during rain and night with little luminance conditions. Details about the three nighttime sequences are described below and summarized in Table 3.2.

Chapter 5 proposes to introduce image blurring method in place recognition under different conditions, including both the daytime and nighttime conditions. Different from the other two proposed approaches in Chapter 4 and Chapter 6, experiment instead of the

Table 3.2: The description of evaluation with day and nighttime sequences in Chapter 5.

| Sequence Condition | Datasets | Purpose | No. of frames | Recorded | Resolution | |
|---|---|---|---|---|---|---|
| | | | | | Original | Resized |
| day | Oxford Robotcar | reference | 236 | 2014/11/18 | 1280×960 | 224×224 |
| night | Oxford RobotCar | query | 236 | 2014/12/16 | 1280×960 | 224×224 |
| night-rain | Oxford RobotCar | query | 236 | 2014/12/17 | 1280×960 | 224×224 |
| night-darker | Oxford RobotCar | query | 236 | 2014/11/14 | 1280×960 | 224×224 |

novel algorithm approach is the most important part. Thus, more sequences are needed to do the experiment. Therefore, the evaluation dataset in Chapter 6 includes both the daytime sequences in Table 3.1 and the nighttime sequences in Table 3.2.

### 3.2.2/ DATA PRE-PROCESSING

Due to different image sizes in different datasets, the images need to be scaled to a 224×224 size for using NIN_ImageNet to get deep features. In fact, the raw sequence of each dataset contains thousands of images. Instead of using all the images, we need to make a selection and choose hundreds of them.

As the two traversals in each sequence represent the same route, which means the robot car passes by the route twice and captures the street images in a constant frequent, each street view can be represented by two images in theory. The selecting process is to make sure that each image in the query traversal must have its corresponding image in the other reference traversal.

The way of selecting the images is as follows: choosing one starting position and one ending position, then the images captured between those two positions are selected and number renamed. Specifically, the image representing the starting position is named as $0$, and after the car moving on about one meter, the camera captures an image at this position and this image is named as $1$. In this way, for example, the car arrives at the ending position, $236$ images composing a sequence as the night sequence in Chapter 6, are selected and named from $1$ to $236$. The same selecting process is used in the two traversals of each sequence. And because the two traversals are taken in the same route, the same starting and the ending positions must be selected from each. Therefore, the same name (number) of the selected images from the two traversals represents the same place.

The distance between two consecutive images in terms of acquisition position changes from 1 to 4 meters. Indeed, the images are not captured based on the car position, but based on time. The speed of the vehicle changes as the traffic conditions change. Therefore, each two neighbouring images which are captured though at the same time lag may lead to different distances. In the image selection process, we tried to make the distance of the neighbouring images as same as possible, but the error change exists as it is a manually labelling process. The proposed place recognition approaches are good in tackling the coarse localization problem. If the distance between two consecutive images is too small, which means the neighbouring two places are very close and have almost the same appearance, the performance of the proposed approaches will decrease.

### 3.2.3/ EVALUATION METRIC AND BASELINE APPROACHES

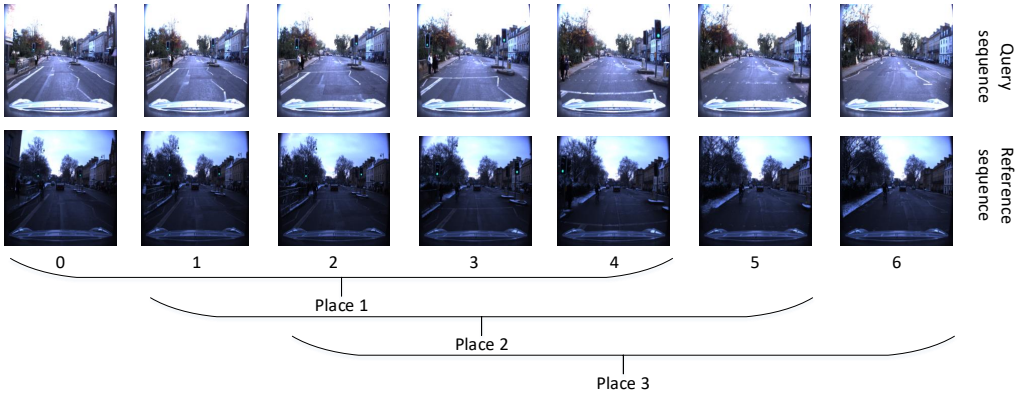



Figure 3.11: The sequence selected from each dataset contains query sequence and reference sequence which represents the same places in different conditions. Neighbour of five images is considered as a tolerance in this thesis. Note that consecutive images are about one meter away.

As described in Fig. 3.1, by comparing the similarity of the query image with each image in the reference sequence, the image which is the most similar to the query image is selected as the best matching. To judge whether this best matching is correct or not, an evaluation metric is needed and will be introduced in the following.

Firstly, we need to define how to decide that two images, query and reference images, correspond to the same place. In this thesis, if two images are within the five neighbouring images (about 5-20 meters of their ground truth positions and therefore achieving a coarse localization), they are considered to be acquired at the same place (see the

process in Fig. 3.11). As the images are named in order, the neighbor five images can be recognized by their name. Then, given the result of the matching process that provides the best matching query and reference images, if those images are within the five neighbouring images, they will be thought as corresponding to the same place and the recognition result is correct.

Unlike the classical object retrieval method [Sivic et al., 2003], the precision rate instead of the recall rate is the evaluation indicator of the place recognition performance. Therefore, the evaluation method in this thesis is just the one used in [Olid et al., 2018], i.e., the best matching reference image is within $5$ meters error tolerance threshold from the query image. And the percentage of the correct matching images to all the query images, called as the precision rate, is the performance criteria in this thesis.

To evaluate the performance of the proposed methods, we compare the results obtained by applying image retrieval on the images resulting from our proposed methods (as seen in Fig. 3.12) with the results obtained by directly applying image retrieval on raw images.

(a) Raw day image (b) Inpainted image (c) Blurred image (d) Raw night image (e) Enhanced image

Figure 3.12: Examples of raw images and their corresponding images resulting from our proposed methods.

The results of the proposed approaches are also compared with other state-of-the-art place recognition approaches: DenseVLAD, OLO [Chen et al., 2017b], and ToDayGAN [Anoosheh et al., 2019].

DenseVLAD is a classical place recognition approach which is seen as a baseline comparison in many papers. OLO uses the CNN network to extract features which is similar to our approach. ToDayGAN tackles the day and night place recognition problem by translating the night-time images into day-time ones firstly and then uses the method in densevlad for the image retrieval process. DenseVLAD is a robust place recognition approach based on the synthesized reference images which have a similar viewpoint with the query image, though the reference images in this paper thesis do not need to change

their viewpoints. Therefore, DenseVLAD and ToDayGAN have the same place recognition process, while ToDayGAN using the translated daytime images instead of the raw images using in DenseVLAD. The connection of these two methods can well evaluate the performance of the translated daytime images by the image-to-image translation network.

## 3.3/ CONCLUSION

In this chapter, the framework of the place recognition which will be used in the contribution works is introduced firstly. Then the experiments setup for the place recognition approaches including evaluation datasets, data pre-processing, evaluation metrics and the compared approaches are described. In the following chapters, the proposed place recognition methods are mainly explained and the overall experiment setup will not be detailed further.

# 4

# VISUAL PLACE RECOGNITION WITH DYNAMIC OBJECTS REMOVAL STEP

In this chapter, a dynamic objects removal step is integrated in the place recognition framework. The aim is to remove the unstable features from an image, thus improving the robustness of the image features and the place recognition performance. This method is achieved by adding two parts ahead the place recognition framework presented in previous Chapter 3: 1) dynamic objects detection and removing using semantic segmentation, 2) background reconstruction using image inpainting. In the experimental part, five daytime sequences are selected from three publicly available datasets (CMU, St. Lucia and Oxford RobotCar) to evaluate the proposed approach. The results are also compared with two state-of-the-art place recognition methods: DenseVLAD [Torii et al., 2015] and OLO [Chen et al., 2017b].

## 4.1/ INTRODUCTION

In the appearance-based solutions for solving the place recognition problem, translating the image appearance into distinguishable features is a crucial step for the success of recognizing the correct place. Therefore, generating distinguishable features to represent images under the challenge of appearance variations determines the results of place recognition in changing environments.

Traditional methods tried to design local features that are viewpoint-, illumination- and scale-invariant, such as speeded-up robust features (SURF) [Bay et al., 2006b] and

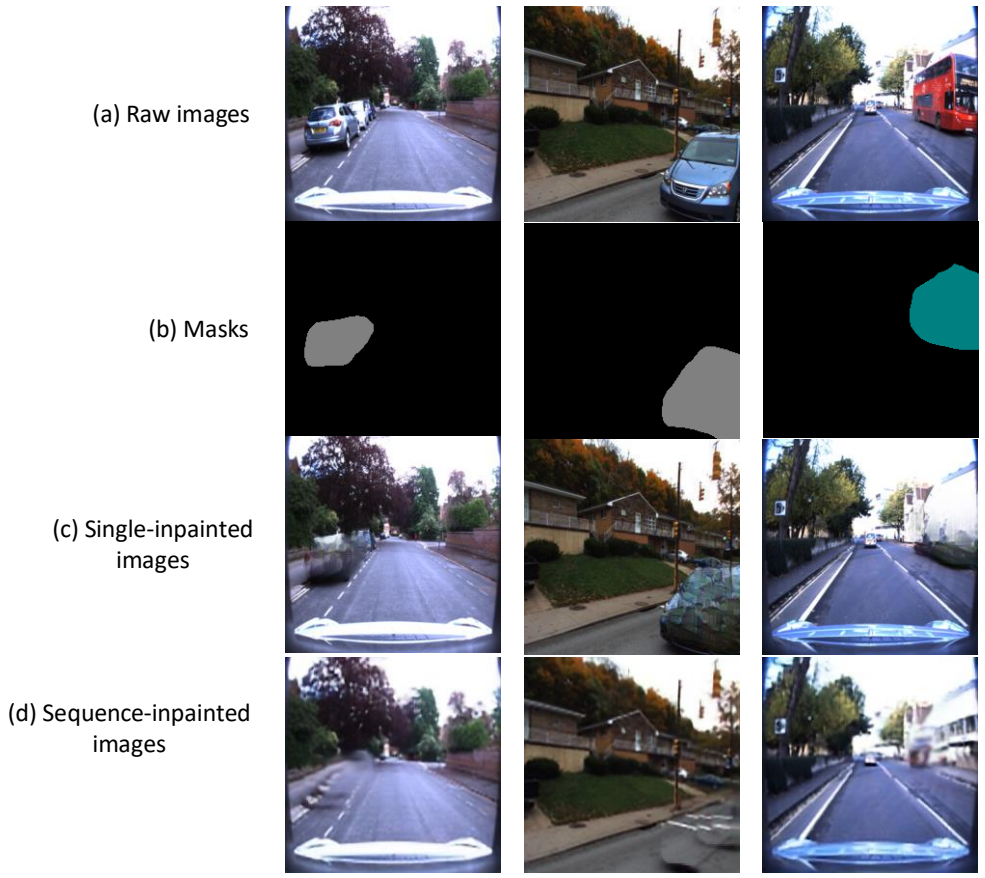



Figure 4.1: Raw images (fist row). Masks (second row) generating by FCN based semantic segmentation. Regions depicted in color are dynamic objects recognized by FCN and are seen as missing regions by inpainting methods. Inpainted images (third row) obtained with single image inpainting method (EdgeConnect). Inpainted images (last row) obtained with sequence image inpainting method.

scale-invariant feature transforms (SIFT) [Lowe, 1999]. In the spirit of recent advances in Convolutional Neural Networks (CNNs), several global features trained by pretrained or learned networks have been applied to place recognition [Bai et al., 2018] [Sünderhauf et al., 2015a].

These global deep features are more efficient than traditional hand-craft local features for place recognition, but they are still not robust enough when facing to severe challenges in real-world place recognition. In reality, the environment in the avenues is very complex. Effectively, in addition to dramatic changes in weather conditions, illumination and viewpoint, dynamic objects, such as vehicles and pedestrians, also contribute to the appearance changes and make it difficult to distinguish the same place at different times.

By carefully analyzing the street views, it can be seen that the stable objects such as buildings, roads, and trees can always be considered as landmark features in place recognition, while it is common for them to be occluded by dynamic objects such as vehicles and pedestrians in the realistic images with huge traffic variation. Before generating the global deep features from an avenue image, selecting the dynamic objects in the image and translate them into a stable background is an idea worth trying. This chapter achieves this idea and proves that deep features from dynamic objects damage the performance of place recognition and proposes a feature selection system to weaken this kind of impact on place recognition.

Benefiting from the development of semantic segmentation and image inpainting approaches, the feature selection system is completed by two steps (Fig. 4.1): 1) dynamic objects detection and removing by semantic segmentation 2) background reconstructing by image inpainting. Considering the environment in the street, the dynamic objects are classified as cars, buses, trucks, and pedestrians. They are firstly recognized by FCN (Fully Convolutional Networks) [Long et al., 2015] and labeled into masks. Then, these labeled masks, which are then seen as missing regions, are filled with fine details using both single-image and sequence-image based inpainting approaches. The outputs of single-image based inpainting method (EdgeConnect [Nazeri et al., 2019]) and sequence-based inpainting method (DVL - deep video learning [Kim et al., 2019]) are called single-inpainted images and sequence-inpainted images respectively, in this chapter. In this way, the inpainted images can be considered as the result of the transformation process from an original image containing both useful and noisy information to a new image containing only stable information useful for place recognition. This process is detailed in section 4.3 and illustrated in Fig. 4.12.

A CNN network is then used to extract global features from these well-inpainted images. The image retrieval process between the current image and the reference images is finally performed using Euclidean distance as a metric. To our best knowledge, this is the first work that deeply applies FCN and inpainting work into place recognition to get robust features representation.

In this chapter, the source of images for experiments is selected from three publicly available datasets: CMU, St. Lucia and Oxford RobotCar [Cummins et al., 2008]. Five daytime sequences of different routes are chosen in the three datasets and their inpainted

images are generated for each sequence. Then, four CNN models (VGG19, NIN_ImageNet, bvlc_GoogLeNet and AlexNet) are used to learn the features of both inpainted images and raw images (raw images correspond to sequences without dynamic objects detection and inpainting), respectively. In each experiment, the results obtained using CNN features of raw-images are compared with the results obtained using CNN features of inpainted images: a noticeable improvement is performed on the task of place recognition by using the CNN features of inpainted-images. At last, the results obtained by our method are compared with two state-of-art place recognition methods: DenseVLAD [Torii et al., 2015] and OLO [Chen et al., 2017b].

The major contributions of this chapter are summarized below:

**1)** Combining semantic segmentation and image inpainting techniques with place recognition method,

**2)** Proposing a feature extraction system that is robust to traffic condition variation, requiring no specific training.

The rest of this chapter is organized as follows: section 4.2 briefly introduces the related methods used in this chapter (semantic segmentation for non-permanent obstacles detection and image inpainting for background reconstruction). The proposed method is then detailed in section 4.3. The experimental setup and results are presented in section 4.4. At the end of this chapter, conclusion and future work are outlined in section 4.5.

## 4.2/ USED METHODS

Semantic segmentation and image inpainting are used to detect and remove the cars and pedestrians from images. This section briefly introduces the methods of these two fields.

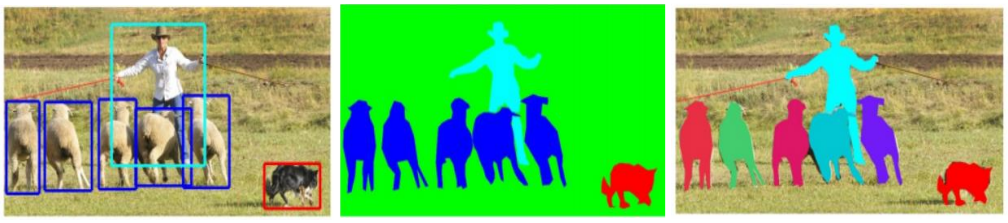
### 4.2.1/ SEMANTIC SEGMENTATION

Computer vision includes many tasks, such as image classification, object detection, semantic segmentation, instance segmentation and panoramic segmentation.

Semantic segmentation is a very important field in computer vision. It refers to the pixel-level recognition of images, that is, to mark the object category to which each pixel in

the image belongs. Different from the object detection which gives bounding boxes as outputs, the classification output of semantic segmentation is more accurate. Fig. 4.2 shows the difference of the outputs of object detection and semantic segmentation.

Instance segmentation is a combination of object detection and semantic segmentation. Compared to object detection, instance segmentation does not need to mark each pixel, but can be more accurate to detect the edge of the object than the bounding box of object detection. Relative to semantic segmentation, instance segmentation needs to distinguish instances of the same category. As shown in Fig. 4.2 (c), instance segmentation use different color outlines to distinguish each sheep.



(a) Object detection (b) Semantic segmentation (c) Instance segmentation

Figure 4.2: Example of object detection, semantic segmentation and instance segmentation.

The categories Proposedof cars and pedestrians instead of the specific individual of the same object are needed to be recognized in our method, so this chapter mainly focuses on semantic segmentation methods.

The common applications of semantic segmentation include:

**1)** Self-driving cars. The real-time segmentation of road scenes during automatic driving can help cars to have better perception of the environment, so that to guarantee the safety of self-driving.

**2)** Medical image diagnosis. Images with segmentation can improve the analysis efficiency of the radiologists. For example, Fig. 4.3 shows the segmentation of chest X-rays, heart (red), lungs (green) and clavicle (blue).

Before applying deep learning (DL) to computer vision, TextonForest [Shotton et al., 2009] and random forest [Fröhlich et al., 2013] were used for semantic segmentation. With the development of deep learning, CNN not only applies

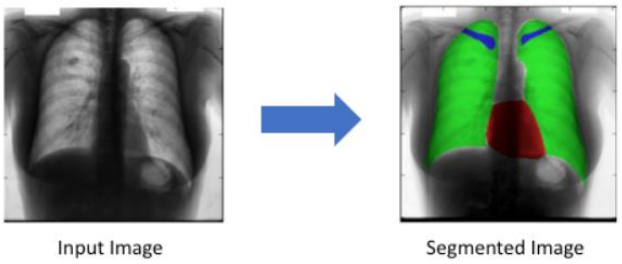



Figure 4.3: Applying semantic segmentation in medical image diagnosis.

in image recognition, but also plays an important role in semantic segmentation. The semantic segmentation methods based on convolutional neural network are different from the traditional semantic segmentation methods as they can automatically learn the image features and greatly improve the accuracy of semantic segmentation.

In the fully convolutional based models, symmetrical semantic segmentation models [Long et al., 2015] [Badrinarayanan et al., 2017] [Ronneberger et al., 2015b] are firstly designed for image segmentation.

FCN [Long et al., 2015] is an early but important work for solving the semantic segmentation problem. It is a fully convolutional network without any fully connected layer to classify images at the pixel level.

As shown in Fig. 4.4, FCN uses image classification networks (such as AlexNet, which was detailed in section 3.1.2) as an encoder. In the traditional CNN architecture, the first 5 layers are convolutional layers, the sixth and seventh layers are respectively a one-dimensional vector with a length of 4096, and the eighth layer is a one-dimensional vector with a length of 1000, the output of the last layer is the possibilities of 1000 different categories. FCN represents these 3 layers as convolutional layers, and the size of the convolution kernels (number of channels, width, height) is (4096,1,1), (4096,1,1), (1000,1,1), respectively. Therefore the output of the last fully convolutional layer is 1000 heatmaps which represent the high-dimensional image features instead of the possibilities of 1000 different categories. As all layers in the FCN network are convolutional layers, it is called fully convolutional network.

After obtaining the heatmap, the most important and final step is upsampling the original image to get the well segmented image as shown in Fig. 4.5. In order to classify and pre-

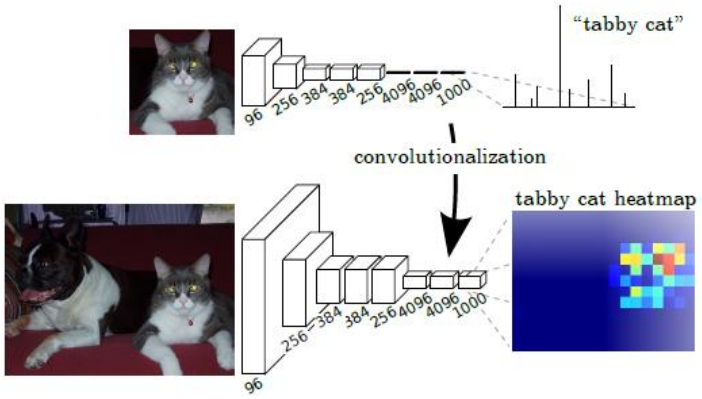



Figure 4.4: FCN uses image classification networks (such as AlexNet) as an encoder and converts its last three fully connected layers into convolutional layers [Long et al., 2015].

dict the label of each pixel, the maximum numerical description (probability) of the same position in 21 heatmap images (last convolutional layer has a kernel of (21, 1, 1) here) is obtained as the classification of this pixel. Because of classifying pixel by pixel, FCN does not well consider the relationship between pixels. It ignores the spatial regularization step used in common pixel-based segmentation method, thus lacking spatial consistency.

After many times of convolution and pooling, the size of the heatmap is smaller and the resolution is lower than the input image. Even if methods like skip connections are used to improve the image resolution, the results of FCN are still not detailed enough, thus not sensitive to the details in the image.

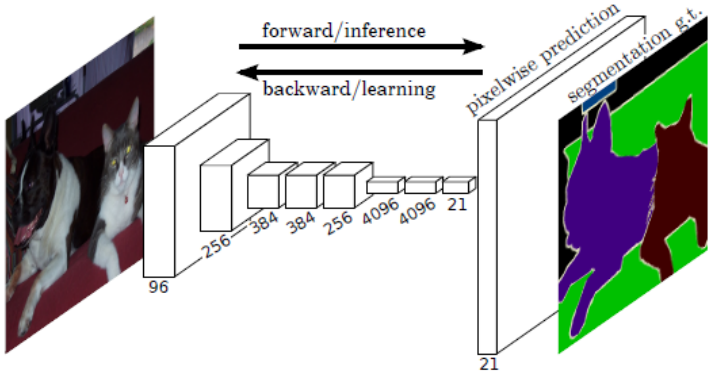



Figure 4.5: Fully convolutional networks [Long et al., 2015].

SegNet [Badrinarayanan et al., 2017] is proposed to solve the problem of losing the de-

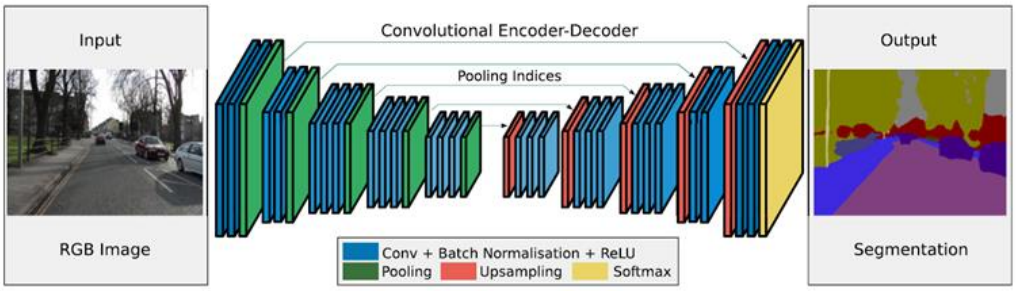



Figure 4.6: The architecture of SegNet [Badrinarayanan et al., 2017].

tails caused by FCN. The design of SegNet and FCN are similar. As it is shown in Fig. 4.6, the encoding part of SegNet is mainly composed of the first 13 convolutional layers and 5 pooling layers of the VGG16 network (introduced in section 3.1.2). The decoding part is also composed of 13 convolutional layers and 5 upsampling layers. The high-dimensional features generated by the decoder's last layer are sent to a trainable softmax classifier to classify each individual pixel. In particular, the SegNet network uses pooling indices to save the contour information of the image, reducing the amount of parameters.

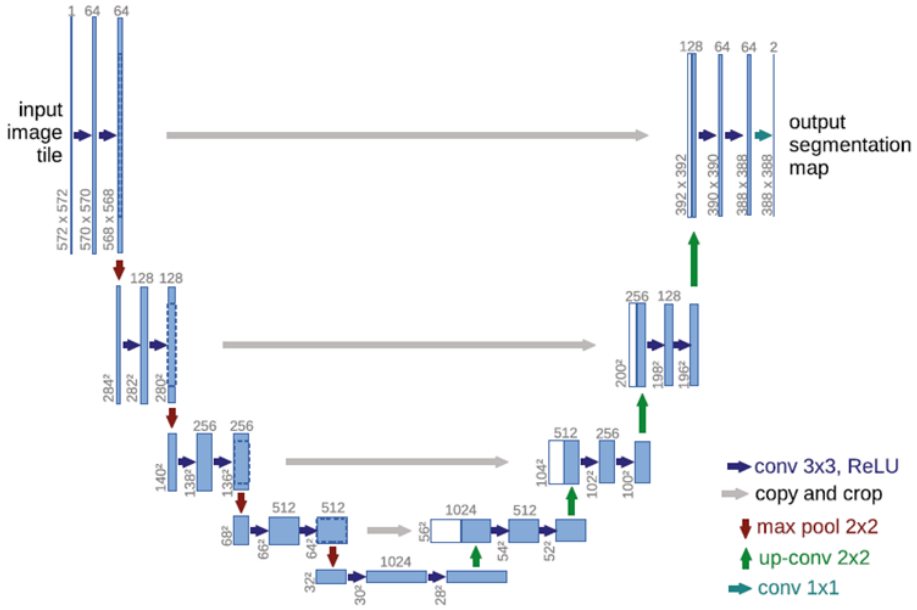



Figure 4.7: U-net architecture [Ronneberger et al., 2015a].

U-net [Ronneberger et al., 2015b] is a symmetric convolutional network designed for

biomedical image segmentation. As shown in Fig. 4.7, the left side of this architecture is the encoder side, used for feature extraction, and the right side is the decoder side, used for upsampling. To learn the network robustness and invariance, data augmentation is applied in the training process to extend the data. U-net achieves a good segmentation effect, but can only process 2D images.

Though these symmetrical semantic segmentation models are chosen in our work of this chapter, other fully convolutional based models for semantic segmentation have also state-of-art performance, such as the extended convolution semantic segmentation model [Chen et al., 2015b] [Lin et al., 2017], residual network semantic segmentation models [Zhao et al., 2017] and GAN semantic segmentation models [Luc et al., 2016].

### 4.2.2/ IMAGE INPAINTING

Image inpainting is a technique that aims to restore a damaged image in an undetectable form. As shown in Fig. 4.8 , the distorted images [ElHarrouss et al., 2019] can be caused by many kinds of distortions, such as text, blocks, noise, scratch, lines or many types of masks [Ružić et al., 2015] [Jin et al., 2015] [Kawai et al., 2016].

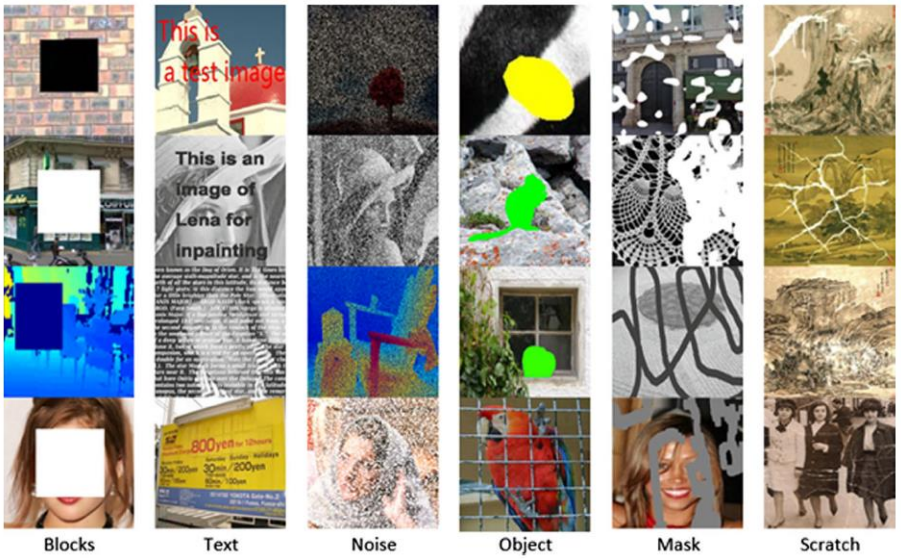



Figure 4.8: Types of distorted images [ElHarrouss et al., 2019].

The inpainting results can be changed from the perfect restoration of the distorted objects to the removal of the objects, or even the unpredictable replacement of the original

objects, thus different kinds of image inpainting applications are developed. For example, image restoration, photo-editing and image coding and transmission are three different directions of image inpainting (detailed in Fig. 4.9). Image restoration aims to remove the scratch and text in the images. Photo-editing can be used to remove any of the objects that are not wanted in the image. And image coding and transmission focus on recovering the missing blocks that are caused by the network packet loss during image transmission. In the research of this chapter, removing the specific objects that are not wanted in the image is the goal we focus on, i.e, the method of photo-editing is needed in this chapter.

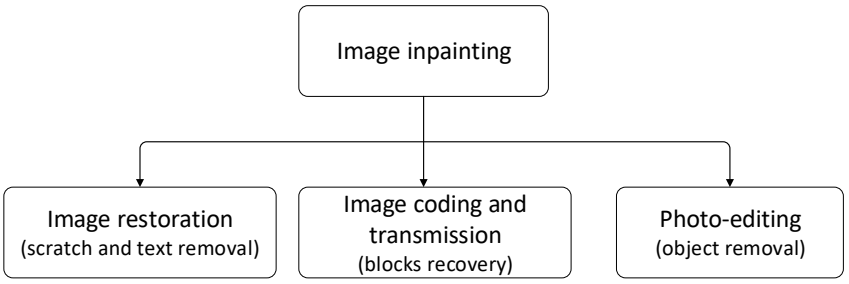



Figure 4.9: Image inpainting applications.

Bertalmio et al [Bertalmio et al., 2000] proposed the concept of inpainting for the first time. Since then, image inpainting has been widely studied and developed. Based on [ElHarrouss et al., 2019], the research of image inpainting includes sequential-based, CNN-based and GAN-based approaches.

The patch-based and diffusion-based inpainting are two methods included in sequential-based image inpainting approach. The key point of patch-based method [Jin et al., 2015] [Kawai et al., 2016] [Guo et al., 2018] [Xue et al., 2017a] [Liu et al., 2018] [Ding et al., 2019] is to select the similar patches in the original images and put them in the regions waiting to be repaired. For the diffusion-based inpainting method [Li et al., 2017], the pixels at the edge of the missing regions are needed to repair in-grow in accordance with the nature of the normal image area, and diffuse to fill the entire missing regions.

The sequential-based approaches can perform well when facing local distortions such as texture damage of images, but are insufficient to repair the images with global information distortions. Therefore, CNN-based approaches

[Weerasekera et al., 2018] [Zhao et al., 2018] [Chang et al., 2019] [Zhu et al., 2018] [Zeng et al., 2019] [Liu et al., 2019] for image inpainting are developed to solve the problem. The encoder-decoder networks are the common framework in CNN-based approaches, such as the encoder-decoder network proposed in [Liu et al., 2019] (as shown in Fig. 4.10). In refinement network, a CSA (coherent semantic attention) layer at the resolution of 32×32 is proposed and applied in.

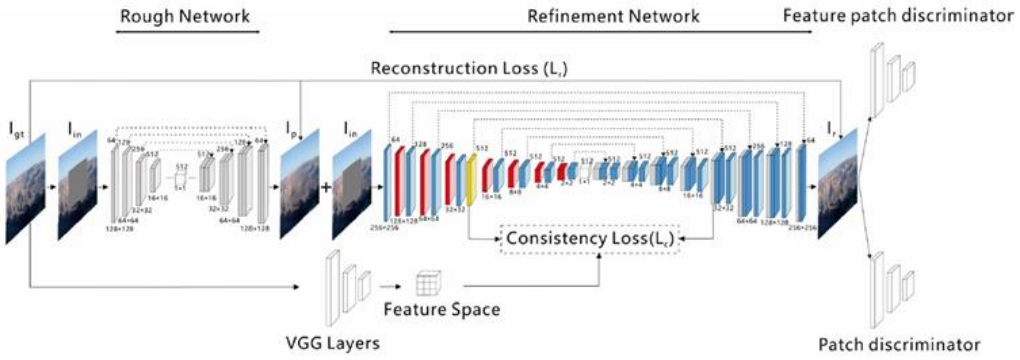



Figure 4.10: Encoder-decoder networks model in [Liu et al., 2019].

As the development of the GAN, more and more GAN-based approaches are proposed to solve the image inpainting problem. Although this is a new research field that needs to be more developed, [Han et al., 2019] [Shin et al., 2021] [Dong et al., 2019] [Nazeri et al., 2019] are the works which have the state-of-the-art performance.

The development of any computer vision direction is inseparable from the support of data, especially in the area of deep learning. Examples of some common image inpainting datasets for different scenes are shown in Fig. 4.11.

**Paris StreetView** [Doersch et al., 2015] comes from Google StreetView which consists of a large-scale street images.

**Places** [Zhou et al., 2018] contains 10 million images including many scene categories, such as streets, bedrooms, canyon, etc.

**Depth image** [Xue et al., 2017b] is a combination of RGB-D and grayscale depth images.

**Foreground-aware** [Xiong et al., 2019] is special for containing masks that can be put in the images for being restored by the image inpainting methods.

**Berkeley segmentation** [Martin et al., 2001] is composed of two kinds of images: RGB and Grayscale images.

**ImageNet** [Russakovsky et al., 2015] consists of thousand of images, most of whose are annotated with a bounding box.

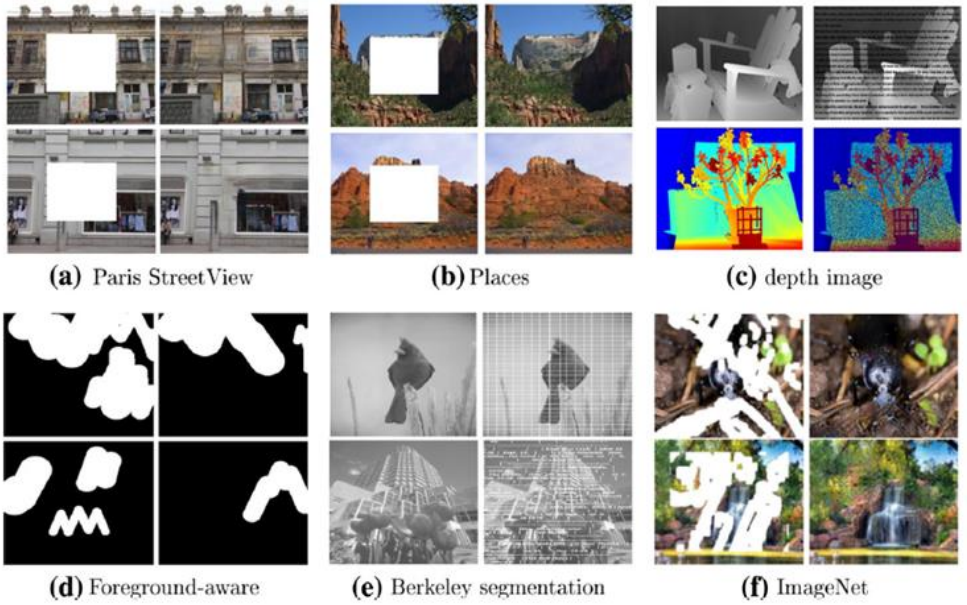



Figure 4.11: Examples from Image inpainting datasets [ElHarrouss et al., 2019].

## 4.3/ THE PROPOSED APPROACH

Our approach in this chapter proposes a robust global deep feature that selects the useless dynamic information, like cars, vehicles, pedestrians, and translates them into useful information by reconstructing the background from the previous images. This is achieved in two stages (as shown in Fig. 4.12): 1) useless objects detection and removing, and 2) image inpainting and reconstruction.

By the two-stages method, each image in the original sequence which consists of both query frames and reference frames is transformed into a new image, called inpaint-image. After this process, each original image will get its inpaint-image. Then the place recognition step (as shown in Fig. 3.1) will recognize the query image from the reference images based on these inpaint-images.

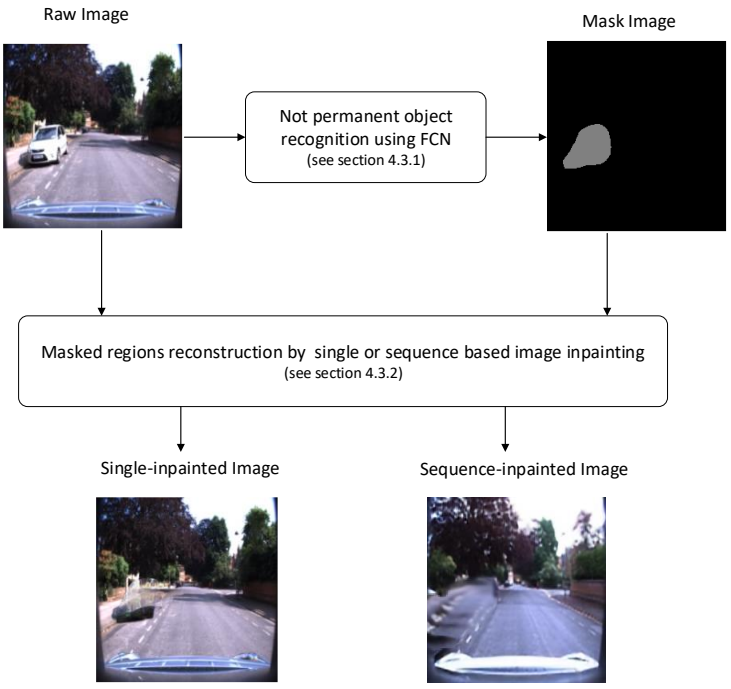



Figure 4.12: Proposed system to decrease the impact of dynamic information of the image and reconstruction of its background.

### 4.3.1/ PROPOSED DYNAMIC OBJECTS REMOVING

The first step of the proposed approach is to eliminate the objects of the image that are not permanent in the scene (as vehicles, pedestrians, $\cdots$). For that, FCN (Fully Convolutional Networks) [Long et al., 2015] is used to recognize these objects in the image and make local predictions of these objects. Given an image of any size as input, this end-to-end fully convolutional network (FCN) provides an output image of the same size with a dense prediction of each pixel (as the mask image in Fig. 4.12). The dynamic objects are not directly removed from the image but are seen as missing regions in the image inpainting methods that are chosen. The section 4.3.2 part will explain this step in detail.

To get the output of FCN, the first step is to choose a fine-tuned classification model as the encoder (details can be found in section 4.2.1), the CNN model -resnet101- which is trained on PASCAL VOC dataset is picked in this chapter. Then the fully connected layer of this CNN model is transformed into convolution layers that are able to make a dense prediction and output a heatmap.

A good semantic segmentation result for vehicles in the road is achieved, though it can fail to recognize pedestrians and far-away objects. This is because of the disadvantage of FCN: lacking sensitivity to details. Also, the method is very convenient as it can use any size of image as input and is time-efficient. That is why it is FCN being chosen in this chapter among all the semantic segmentation methods listed in section 4.2.1.

The output image of FCN will be considered as an input mask image in image inpainting step.

Given a raw image and its mask image generated by FCN, the output of EdgeConnect is our expected inpainting-image. In the process of inpainting, EdgeConnect will see the dark part of the mask image as a background and recognize the remaining parts as missing regions which have to be reconstructed.

### 4.3.2/ IMAGES REPAIRING WITH IMAGE INPAINTING APPROACH

Image inpainting means reproducing the missing regions of a ground truth image as if it has not been corrupted. In our thesis, we tested a single image inpainting method as well as an image sequence inpainting method to repair the background information of an image based on the previous mask image generated by FCN. The single image inpainting method reconstructs an image based on the current image itself only, and can be done online. The image sequence inpainting method reconstructs an image based on not only the current image, but also the previous and the future images, thus it has to be done off-line.

#### 4.3.2.1/ SINGLE IMAGE INPAINTING - EDGECONNECT

EdgeConnect [Nazeri et al., 2019] is a single-image based inpainting method that trains an edge generator to generate the hallucinate edges of the missing regions in the first step, then reconstructs the missing regions by learning the hallucinate edges. Therefore it consists of two GAN (Generative Adversarial Network) models. The first model generates the edges of an image, and the second completes the texture and color information.

The model of GAN is an adversarial process of two networks, generator, and discriminator. The generator network should generate fake data from noise and make it as real as

possible, and the discriminator network should try to distinguish the generated fake data from the real data. Training a GAN is like to find the optimal point in the minimax game. It can be represented as:

$$\begin{aligned}\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}\left[\log D(x)\right] \\ + \mathbb{E}_{z\sim P_z(z)}[\log\left[(1-D(G(z)))\right],\end{aligned} \tag{4.1}$$

where $z$ is the noisy vector with its distribution $p_z(z)$ and $G(z)$ represents the fake information generated by the generator network. $x$ is the training sample whose distribution is trained as close to $P_{data}$, $D(x)$ represents classifying $x$ as true or false by the discriminator network. $\mathbb{E}$ is used to denote the expected value.

The architecture of generators in EdgeConnect is based on the network proposed by [Johnson et al., 2016a]. The discriminators in EdgeConnect follow the $70 \times 70$ PatchGAN [Isola et al., 2017] architecture, which gives an $N \times N$ array of outputs instead of a single scalar output given by the regular GAN discriminator. The difference is that the $N \times N$ array of outputs determines real or fake each of the overlapping image patches, while the single scalar output determines whether the whole image is real or not. For all the layers of the network, instance normalization (for each sample, subtract mean and divide by standard deviation) is used.

The details of the two GAN models of EdgeConnect, edge generator and image completion network, are as follows (as shown in Fig. 4.13).

1) Edge Generator

Given the ground truth image $\mathbf{I}_{gt}$ (i.e the raw image of our dataset), $\mathbf{C}_{gt}$ and $\mathbf{I}_{gray}$ are its edge map and grayscale format image, respectively. Within the image mask $\mathbf{M}$ (i.e. the mask in Fig. 4.13), the masked grayscale image $\tilde{\mathbf{I}}_{gray} = \mathbf{I}_{gray} \odot (\mathbf{1} - \mathbf{M})$ (i.e. the input grayscale in Fig. 4.13) and the corresponding masked edge map $\tilde{\mathbf{C}}_{gt} = \mathbf{C}_{gt} \odot (\mathbf{1} - \mathbf{M})$ (i.e. the input edge image in Fig. 4.13) can be easily described. Here, $\odot$ means the Hadamard product.

Then the generator $G_1$ uses the masked grayscale image $\tilde{\mathbf{I}}_{gray}$ and the masked edge map $\tilde{\mathbf{C}}_{gt}$ as inputs and generates a new edge map $\mathbf{C}_{pred}$ (cf. Eq. (4.2), example image can be seen in Fig. 4.13) which hallucinates edges in the masked regions.

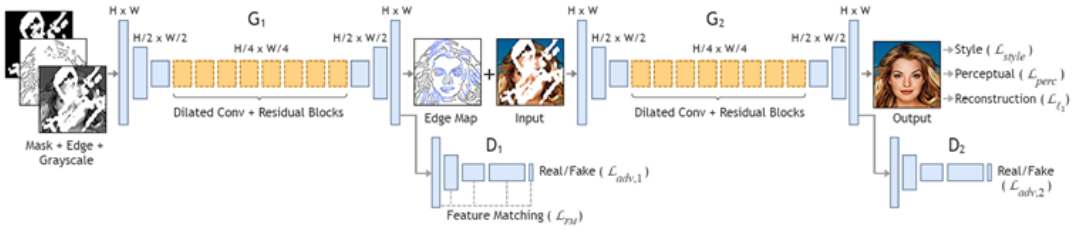



Figure 4.13: Summary of the single image inpainting method - EdgeConnect [Nazeri et al., 2019]. Taking raw image and its mask as input, the edge generator ($G_1$ as the generator and $D_1$ as the discriminator) generates the predicted edge map. Based on the predicted edge map and the input images, the image completion network ($G_2$ as the generator and $D_2$ as the discriminator) completes the inpainting task and gives the inpainted image as a final output.

$$\mathbf{C}_{pred} = G_1\left(\tilde{\mathbf{I}}_{gray}, \tilde{\mathbf{C}}_{gt}, \mathbf{M}\right). \tag{4.2}$$

To determine whether the generated edge map is real or not, the discriminator $D_1$ uses $\mathbf{C}_{gt}$ and $\mathbf{C}_{pred}$ conditioned on $\mathbf{I}_{gray}$ as inputs. The loss function to train the network combines the adversarial loss and the feature-matching loss and is defined as follows:

$$\min_{G_1}\max_{D_1}\mathcal{L}_{G_1} = \min_{G_1}\left(\lambda_{adv,1}\max_{D_1}(\mathcal{L}_{adv,1}) + \lambda_{FM}\mathcal{L}_{FM}\right) \tag{4.3}$$

Here, $\lambda_{adv,1}$ and $\lambda_{FM}$ are regularization parameters. The parameter setup is not changed in this work, so $\lambda_{adv,1} = 1$ and $\lambda_{FM} = 10$.

The adversarial loss $\mathcal{L}_{adv,1}$ is necessary for training the overall GAN structure and is as follows:

$$\begin{aligned}\mathcal{L}_{adv,1} = {} & \mathbb{E}_{(\mathbf{C}_{gt},\mathbf{I}_{gray})}\left[\log D_1\left(\mathbf{C}_{gt},\mathbf{I}_{gray}\right)\right] \\ & + \mathbb{E}_{\mathbf{I}_{gray}}\log\left[1 - D_1\left(\mathbf{C}_{pred},\mathbf{I}_{gray}\right)\right],\end{aligned} \tag{4.4}$$

$\mathcal{L}_{FM}$ is the feature-matching loss [Wang et al., 2018]:

$$\mathcal{L}_{FM} = \mathbb{E}\left[\sum_{i=1}^{L}\frac{1}{N_i}\|D_1^{(i)}(\mathbf{C}_{gt}) - D_1^{(i)}(\mathbf{C}_{pred})\|_1\right]. \tag{4.5}$$

In this function, $L$ is the last convolution layer of the discriminator, $N_i$ is the number of elements in the $i^{th}$ activation layer, and $D_1^{(i)}$ is the activation in the $i^{th}$ layer of the discrimi-

nator [Nazeri et al., 2019]. This loss function requires the features of the generated edge image $\mathbf{C}_{pred}$ and the real edge image $\mathbf{C}_{gt}$ to be similar, thus making the training process more stable and improving the quality of the images generated by the generator $G_1$.

2) Image Completion Network

Having a masked color image $\tilde{\mathbf{I}}_{gt} = \mathbf{I}_{gt} \odot (\mathbf{1} - \mathbf{M})$ as input, image completion network should first generate a complete image with the condition of a composite edge map $\mathbf{C}_{comp} = \mathbf{C}_{gt} \odot (\mathbf{1} - \mathbf{M}) + \mathbf{C}_{pred} \odot \mathbf{M}$. This generated complete image keeps the original resolution of the input masked image and is described as:

$$\mathbf{I}_{pred} = G_2 \left( \tilde{\mathbf{I}}_{gt}, \mathbf{C}_{comp} \right). \tag{4.6}$$

While the generated image $\mathbf{I}_{pred}$ should try to be the same as the real image $\mathbf{I}_{gt}$, the discriminator in this network should also be trained to recognize that the generated image is false. To achieve this, a joint loss comprised of $\ell_1$ loss, adversarial loss, perceptual loss, and style loss is needed to train the network:

$$\mathcal{L}_{G_2} = \lambda_{\ell_1} \mathcal{L}_{\ell_1} + \lambda_{adv,2} \mathcal{L}_{adv,2} + \lambda_p \mathcal{L}_{perc} + \lambda_s \mathcal{L}_{style}. \tag{4.7}$$

We setup the parameters as in [Nazeri et al., 2019], $\lambda_{\ell_1} = 1$, $\lambda_{adv,2} = \lambda_p = 0.1$ and $\lambda_s = 250$.

For the loss function, $\mathcal{L}_{\ell_1}$ is the $\ell_1$ loss between the real image $\mathbf{I}_{gt}$ and the generated image $\mathbf{I}_{pred}$. It is normalized by the mask size to guarantee the proper scaling.

The adversarial loss is defined as below:

$$\begin{aligned} \mathcal{L}_{adv,2} = {} & \mathbb{E}_{(\mathbf{I}_{gt}, \mathbf{C}_{comp})} \left[ \log D_2 \left( \mathbf{I}_{gt}, \mathbf{C}_{comp} \right) \right] \\ & + \mathbb{E}_{\mathbf{C}_{comp}} \log \left[ 1 - D_2 \left( \mathbf{I}_{pred}, \mathbf{C}_{comp} \right) \right]. \end{aligned} \tag{4.8}$$

The perceptual loss is proposed by [Johnson et al., 2016b] and is defined as:

$$\mathcal{L}_{perc} = \mathbb{E} \left[ \sum_i \frac{1}{N_i} \| \phi_i(\mathbf{I}_{gt}) - \phi_i(\mathbf{I}_{pred}) \|_1 \right]. \tag{4.9}$$

$\phi_i$ is the activation map of the $i^{th}$ layer of a pre-trained network, which in this work is VGG19 network pre-trained on the ImageNet dataset. The layers relu1_1, relu2_1, relu3_1, relu4_1 and relu5_1 of this VGG19 network are chosen to generate activation maps

$\phi_i$. By calculating the $\ell_1$ loss between the generated activation maps of the output image $\phi_i(\mathbf{I}_{pred}$ and of the real image $\phi_i(\mathbf{I}_{gt})$, the details of the feature can be better repaired.

Given the sizes of feature maps $C_j \times H_j \times W_j$, the style loss is expressed as follows:

$$\mathcal{L}_{style} = \mathbb{E}_j \left[ \|G_j^{\phi}(\tilde{\mathbf{I}}_{pred}) - G_j^{\phi}(\tilde{\mathbf{I}}_{gt})\|_1 \right]. \tag{4.10}$$

where $G_j^{\phi}$ is a $C_j \times C_j$ Gram matrix constructed from $\phi_j$. By using the Gram matrix to calculate the difference of the two feature maps, the generator can have a better understanding of the overall image style.

#### 4.3.2.2/ SEQUENCE IMAGE INPAINTING - DVL (DEEP VIDEO LEARNING)

In a sequence of images, the missing parts of an image can be often found in the neighboring images thanks to the changes of viewpoints or the movement of the objects. Therefore, instead of reconstructing the missing parts, borrowing the removed content of the neighboring images is a better solution to recover the current image. To achieve this, the method in [Kim et al., 2019] is used in this chapter.

Given the masked sequence $X_1^T := \{X_1, X_2, \ldots, X_T\}$, the inpainted sequence $\hat{Y}_1^T := \{\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_T\}$ is generated by the inpainting network. The reconstructed regions in this generated sequence should have little difference compared to the ground-truth sequence $Y_1^T := \{Y_1, Y_2, \ldots, Y_T\}$. Then, the conditional distribution $p(\hat{Y}_1^T | X_1^T)$ should be trained identical to $p(Y_1^T | X_1^T)$ and should be represented as:

$$p(\hat{Y}_1^T | X_1^T) = \prod_{t=1}^{T} p(\hat{Y}_t | X_{t-N}^{t+N}, \hat{Y}_{t-1}, M_t). \tag{4.11}$$

where $X_{t-N}^{t+N}$ means the sequence of neighbour images with missing regions needed in multi-to-single frame inpainting, which is achieved by inputting both the original image sequence and its corresponding mask sequence. As in [Kim et al., 2019], $X_{t-N}^{t+N} = X_{t-6}, X_{t-3}, X_t, X_{t+3}, X_{t+6}$. $X_t$ means the current occluded frame. $X_{t-6}, X_{t-3}, X_{t+3}, X_{t+6}$ represent two lagging and two leading frames of the current occluded frame, which are sampled with a stride of 3. $\hat{Y}_{t-1}$ is the previously well-inpainted frame and $M_t$ is a recurrent memory encoding all previous history.

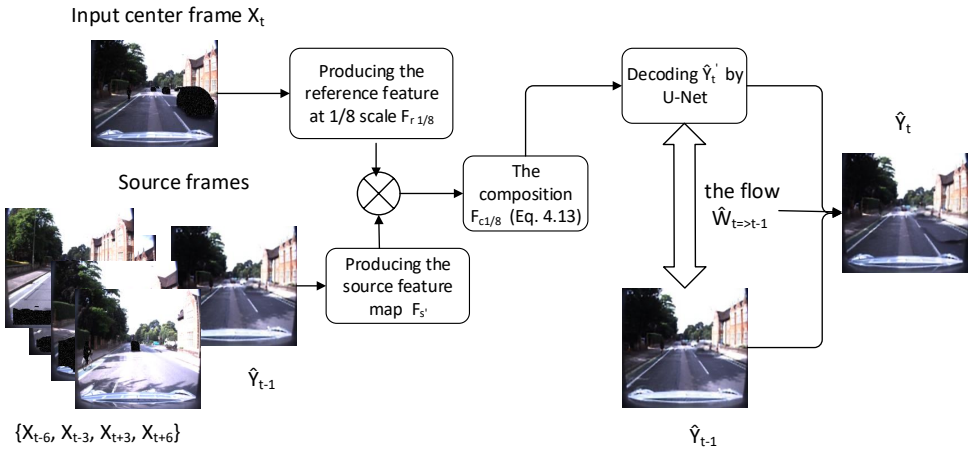



Figure 4.14: Summary of the sequence image inpainting method. The inputs of the network are an image sequence consisting of multiple frames $(X_{t-6}, X_{t-3}), X_t, (X_{t+3}, X_{t+6})$ as well as the previous generated frame $\hat{Y}_{t-1}$. And the outputs are the inpainted frame $\hat{Y}_t$ and $\hat{W}_{t\Rightarrow t-1}$.

Furthermore, a deep network is constructed to model $p(\hat{Y}_t|X_{t-N}^{t+N}, \hat{Y}_{t-1}, M_t)$ as $\hat{Y}_t = D(X_{t-N}^{t+N}, \hat{Y}_{t-1}, M_t)$. This deep network not only learns feature composition of past and future frames, but also applies feature flow learning into dealing the inpainting problem.

1) Learnable feature composition

The inpainting network can simply divide multi-frames into the source and reference streams. The reference stream takes the current frame and its mask as input. For the source stream, two past frames $(X_{t-6}, X_{t-3})$, two future frames $(X_{t+3}, X_{t+6})$ and the previously well-inpainted frame $\hat{Y}_{t-1}$ are provided. The aligned feature maps of these five source streams are concatenated and feed into a $5 \times 3 \times 3$ convolution layer and then an aggregated feature map $F_{s'}$ is produced with the time dimension of 1.

Given the aggregated feature map $F_{s'}$ and the reference feature $F_r$, a mask sub-network is used to regress the composition mask $M$. The input of the mask sub-network is the absolute difference of two feature maps:

$$\Delta F_t = |F_t - F_t^{'}|, \tag{4.12}$$

and a single channel mask $m$ is generated by three convolutional layers with stride one. Here, the composite features $F_c$ can be obtained by a linear combination of $F_r$ and $F_{s'}$ in

four different spatial scales ($1/8, 1/4, 1/2$ and, 1). At the finest scale $1/8$, the composition is calculated by:

$$F_{c_{1/8}} = (1 - m) \odot F_{r_{1/8}} + m \odot F_{s'_{1/8}}, \tag{4.13}$$

where $\odot$ is the element-wise product operator.

From the composite features $F_{c_{1/8}}$, the current raw output $\hat{Y}'_t$ is decoded by U-net [Ronneberger et al., 2015b] which is introduced in section 4.2.1.

2) Feature flow learning

Instead of directly using the current raw output $\hat{Y}'_t$ as the generated image, the inpainting method adopts feature flow learning. The traceable features of the previous output $\hat{Y}_{t-1}$ remain unchanged and the untraceable points are borrowed from the raw output of $\hat{Y}'_t$, which is a composite image from neighboring frames. To achieve this, a flow sub-network is inserted to estimate the flow $\hat{W}_{t \Rightarrow t-1}$ between the previous output $\hat{Y}_{t-1}$ and the current raw output $\hat{Y}'_t$ at the finest scale. By using the estimated optical flow, the previous output $\hat{W}_{t \Rightarrow t-1}$ is warped to get a new output image. And combining this warped image and the raw output $\hat{Y}'_t$ with the composition mask $m_1$, the final output $\hat{Y}_t$ is generated as:

$$\hat{Y}_t = (1 - m_1) \odot \hat{Y}'_t + m_1 \odot \hat{W}_{t \Rightarrow t-1} \hat{Y}_{t-1} \tag{4.14}$$

where $m_1$ is generated by the mask sub-network taking the absolute difference of the two feature maps of $\hat{Y}'_t$ and $\hat{W}_{t \Rightarrow t-1} \hat{Y}_{t-1}$, as the process of generating $m$.

To be summarized, the details of this image sequence inpainting method is shown in Fig. 4.14.

### 4.3.3/ PLACE RECOGNITION METHOD

After generating the inpainted images by one of the methods proposed above, a place recognition step is needed to recognize the reference image that best match with the query image. This process was described in chapter 3.1.

First of all, CNN model is used to generate the deep features of each inpainted image in the well-inpainted query and reference sequences. Then, Euclidean distance is used as

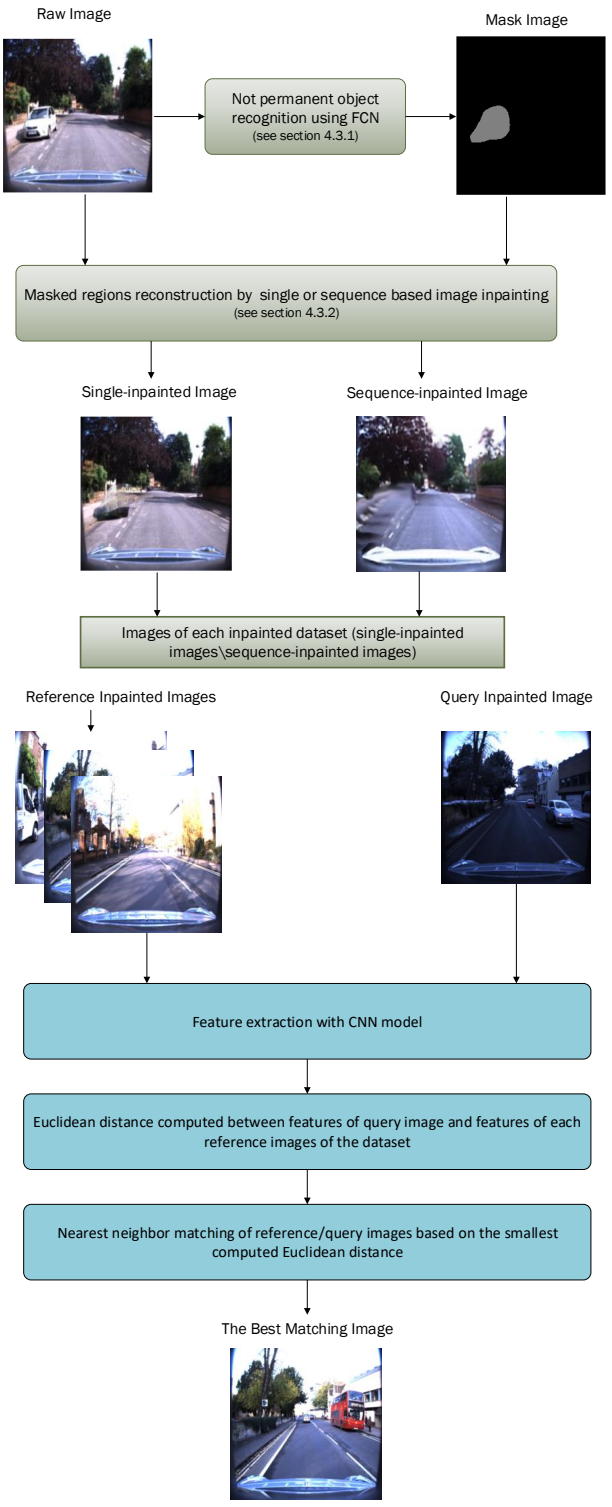



Figure 4.15: Summary of the place recognition method proposed in this chapter. The yellow section is the dynamic objects removal step and the blue section is the place recognition step.

a metric to compute the distance between the features of query image and the features of each reference image. The reference image which has the closest feature distance with the query image is seen as the best matching image.

Fig. 4.15 shows the summary of the method proposed in this chapter. It consists of two steps: dynamic objects removal and place recognition process.

## 4.4/ EXPERIMENTS AND RESULTS ANALYSIS

Experiments in this chapter are developed in three parts.

Firstly, to verify the effectiveness of our approach considering different networks, four kinds of CNN models (AlexNet, VGG19, NIN_ImageNet and bvlc_GoogLeNet) are used to generate the image features. Place recognition results based on each network permit to compare the place recognition accuracy obtained by two state-of-art place recognition methods, and by our method using original images and their corresponding single-inpainted and sequence-inpainted images for three different sequences.

Using our place recognition method without dynamic objects removal process -using the original images- as a baseline for comparison of accuracy of the objects removal based method, how much the object removal method upgrades the performance of place recognition can be proved.

Comparing the results based on different CNN networks is to evaluate the generalization ability of the proposed objects removal methods. This evaluation is helping when applying our objects removal methods in other CNN-based place recognition method.

The dataset used in the experiments are 5 sequences from 3 open source datasets, as detailed in Table 3.1. Each sequence are consists of two types of images -query and reference images- which have the same number and represent the same places at different times. To extend the dataset, the second experiment is designed to test the place recognition performance based on the dataset whose query and reference images are inversed. Fig. 4.16 shows the example images of day5 sequence. In the original sequence of day5, the summer images are set as the query images and the winter images are set as the reference images. After inversing, the winter images are set as the query images and the summer images are set as the reference images. In this way, 5 sequences

of the dataset are extended to 10 without needing of capture new images.

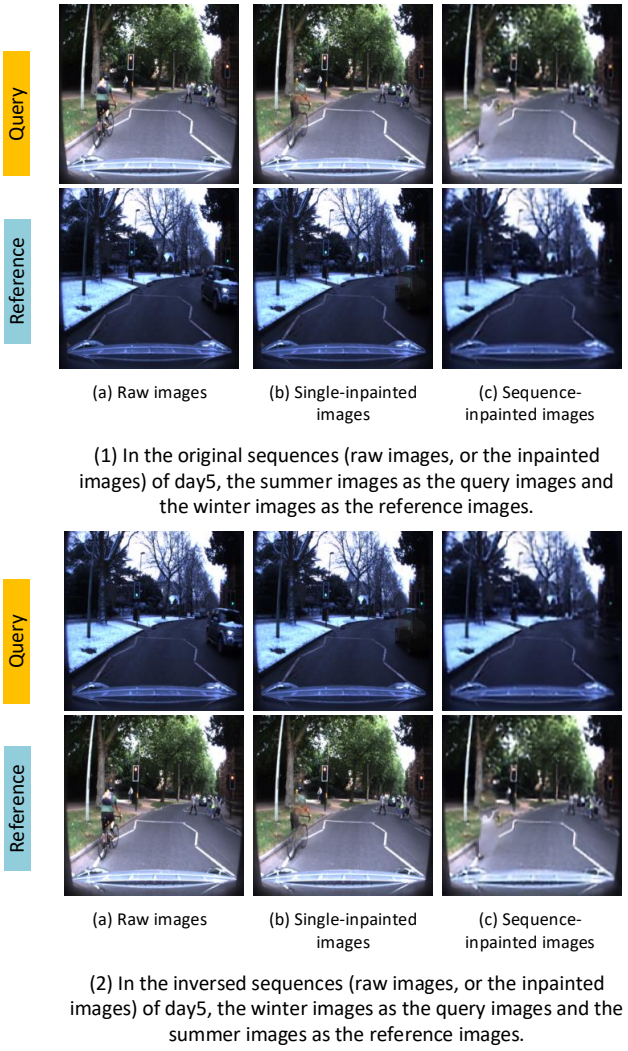



Figure 4.16: Example images of day5 sequence to explain the experiment setup of inversing the query and reference images.

Finally, considering the time efficiency, inpainting step can be only used for reference images or query image. Fig. 4.17 shows some example images of sequence day1 to explain the experiment setup. The setup of Fig. 4.17 (1) using the inpainted query and inpainted reference images is seen as a baseline. Then experiments are done based on the inpainted query and raw reference images (as shown in Fig. 4.17 (2)); the raw query and inpainted reference images (as shown in Fig. 4.17 (3)); The evaluation set shown in Fig. 4.17 (3) has got good results, i.e., query images need not to be inpainted, thus saving the time consuming.

The experiment setup is summarized in Table 4.1.

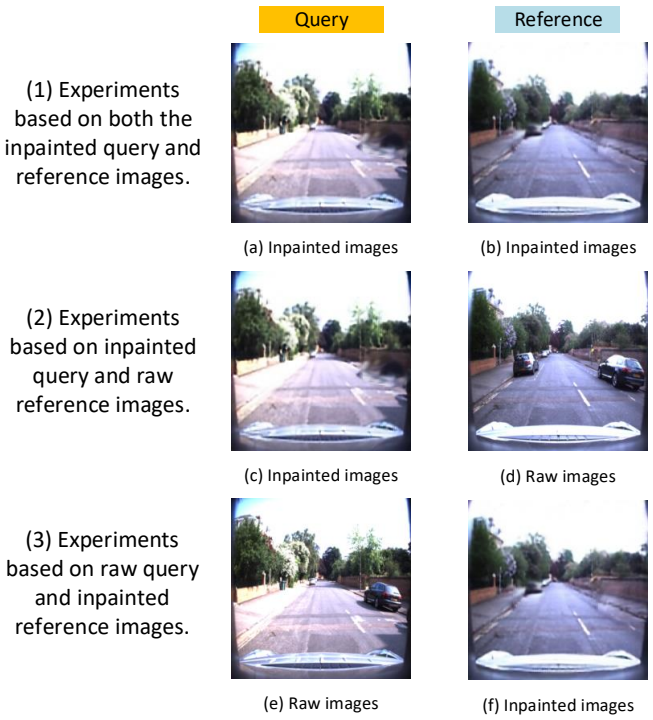



Figure 4.17: Example images of day1 sequence to explain the experiment setup of applying inpainting step only for query or sequence image.

Table 4.1: Summary of the experiment setup.

| Name | Setup | Purpose |
|---|---|---|
| baseline | The proposed place recognition method without dynamic objects removal process, and two place recognition approaches | see how much the dynamic object removal method improves the place recognition performance |
| Section 4.4.1 Setup 1 | Using four pre-trained CNN networks | prove the generalization ability of the proposed dynamic objects removal method for CNN-based place recognition method |
| Section 4.4.2 Setup 2 | inversing the query and reference images of the 5 sequences in the dataset | extend the dataset without capturing new images |
| Section 4.4.3 Setup 3 | considering image inpainting step only for query images and not for reference images;<br>considering image inpainting step only for reference images and not for query images | improve time efficiency for only inpainting query/reference image |

To illustrate these results, the place recognition results are described and analyzed based on each experiment setup in the following parts.

### 4.4.1/ PLACE RECOGNITION RESULTS ACCORDING TO THE USED PRE-TRAINED NETWORKS (CNN MODELS: VGG19, NIN_IMAGENET, BVLC_-GOOGLENET AND ALEXNET)

The two kinds of inpainted images generated in the proposed approach are used as inputs by CNN model to get their feature representation. In this part, four CNN models - VGG19, NIN_ImageNet, bvlc_GoogleNet and AlexNet are considered to extract image features of all the raw images, single-inpainted images and sequence inpainted images from five sequences. Then the place recognition results based on the different feature representations are compared with the state-of-the-art approaches - DenseVLAD [Torii et al., 2015] and OLO [Chen et al., 2017b] for each sequence.

Table 4.2 to Table 4.5 indicate the precision obtained in terms of place recognition (percentage of well retrieved image according to the definition given in chapter3, section 3.2.3) for the proposed approach, and the two approaches of the literature DenseVLAD and OLO, with respect to the considered CNN model (VGG for Table 4.2, NIN_ImageNet for Table 4.3, bvlc_GoogLeNet for Table 4.4, and AlexNet for Table 4.5 respectively). For our approach, three situations are considered: deep features are computed on raw images, or on inpainted images using single-image approach (EdgeConnect) or on inpainted images using sequence DVL approach).

By analyzing these precision results, we can note that the place recognition performance based on the proposed inpainted images are always better than or at least the same as the performance based on the raw images. These are exceptions: the result using single-inpainted images is worse than the result using raw images for the sequences - day1 and day3 - in the case of using AlexNet to generate the features.

For the sequences day1, day2 day3 and day4, the results based on the sequence-inpainted images are better than or once the same as the results based on the single-inpainted images for all the four CNN models used to generate the image features.

Opposite to the other sequences, the best result of sequence day5 is obtained by using single-inpainted images for all the four CNN models. The result obtained by sequence-

inpainted images is better than the result by raw images, but worse than the results by the single-inpainted images.

When compared with DenseVLAD and OLO, the place recognition precision has an obvious improvement using the proposed approaches for the sequence day1, day3 and day5. For the sequences day2 and day4, the proposed approach does not provide a significant advantage with respect to the other approaches, but the results using the sequence-inpainted images are almost similar to the results of DenseVLAD and OLO.

To summarise, depending on the sequences, the ”best” CNN model is not always the same, but the results vary quite little, one can then conclude that the chosen CNN model (among the four tested) for deep feature generation has little influence on the place recognition performance. In this way, the dynamic objects removal method proves its generalization ability as it can be applied in different kinds of CNN-based place recognition approaches.

Table 4.2: The precision comparison of VGG19 based place recognition when using raw images and inpainted-images with two advanced methods.

| Precision(%) | Ours | | | DenseVLAD | OLO |
|---|---|---|---|---|---|
| | raw | single-inpainted | sequence-inpainted | | |
| day1 | 93.02 | 94.57 | 96.51 | 89.15 | 73.64 |
| day2 | 84.54 | 86.96 | 93.24 | 93.24 | 93.24 |
| day3 | 90.50 | 91.50 | 93.50 | 73.00 | 81.50 |
| day4 | 98.24 | 99.56 | 99.56 | 96.92 | 100.00 |
| day5 | 90.78 | 98.58 | 94.68 | 38.30 | 66.67 |

Table 4.3: The precision comparison of NIN_ImageNet based place recognition when using raw images and inpainted-images with two advanced methods.

| Precision(%) | Ours | | | DenseVLAD | OLO |
|---|---|---|---|---|---|
| | raw | single-inpainted | sequence-inpainted | | |
| day1 | 96.90 | 97.29 | 97.67 | 89.15 | 73.64 |
| day2 | 92.75 | 96.14 | 96.62 | 93.24 | 93.24 |
| day3 | 89.00 | 90.00 | 94.50 | 73.00 | 81.50 |
| day4 | 97.80 | 99.12 | 99.12 | 96.92 | 100.00 |
| day5 | 86.88 | 90.43 | 90.07 | 38.30 | 66.67 |

Table 4.4: The precision comparison of bvlc_GoogLeNet based place recognition when using raw images and inpainted-images with two advanced methods.

| Precision(%) | Ours | | | DenseVLAD | OLO |
|---|---|---|---|---|---|
| | raw | single-inpainted | sequence-inpainted | | |
| day1 | 96.12 | 96.90 | 98.06 | 89.15 | 73.64 |
| day2 | 87.44 | 92.27 | 92.27 | 93.24 | 93.24 |
| day3 | 93.00 | 94.00 | 94.50 | 73.00 | 81.50 |
| day4 | 99.56 | 99.56 | 99.56 | 96.92 | 100.00 |
| day5 | 96.10 | 98.23 | 96.45 | 38.30 | 66.67 |

Table 4.5: The precision comparison of AlexNet based place recognition when using raw images and inpainted-images with two advanced methods.

| Precision(%) | Ours | | | DenseVLAD | OLO |
|---|---|---|---|---|---|
| | raw | single-inpainted | sequence-inpainted | | |
| day1 | 97.29 | 96.51 | 97.67 | 89.15 | 73.64 |
| day2 | 87.92 | 92.75 | 96.62 | 93.24 | 93.24 |
| day3 | 94.00 | 93.00 | 95.50 | 73.00 | 81.50 |
| day4 | 98.68 | 99.56 | 99.56 | 96.92 | 100.00 |
| day5 | 88.65 | 91.84 | 89.72 | 38.30 | 66.67 |

Table 4.6: The precision comparison of NIN_ImageNet based place recognition when using raw images and inpainted-images with two advanced methods (but inversing the query and reference images compared to the experiment setup of Table 4.3.)

| Precision(%) | Ours | | | DenseVLAD | OLO |
|---|---|---|---|---|---|
| | raw | single-inpainted | sequence-inpainted | | |
| Dataset with the reversed query and reference sequences | | | | | |
| day1 | 90.70 | 93.41 | 96.90 | 96.51 | 83.33 |
| day2 | 82.61 | 88.41 | 94.20 | 93.72 | 86.47 |
| day3 | 78.50 | 85.50 | 85.00 | 66.50 | 79.00 |
| day4 | 98.24 | 97.36 | 98.24 | 96.92 | 100.00 |
| day5 | 77.66 | 80.14 | 82.27 | 38.30 | 67.38 |

### 4.4.2/ PLACE RECOGNITION RESULTS OBTAINED BY INVERSING THE REFERENCE AND QUERY IMAGES

As introduced in chapter 3, the principle of the place recognition method proposed in this thesis is the following: given a query image, the features of each reference image in the reference sequence are computed and compared with the feature of the query image to identify the reference image that best corresponds to the query image. If the roles of query images dataset and reference images dataset are inverted, i.e. given a reference

image in the original reference image dataset, the feature of each image in the original query image dataset is computed and compared with the feature of the reference image. In this way, the five sequences in our dataset can be extended to 10 sequences without adding new images.

Therefore, the experiment carried out in this part which inverses the query and reference images in each sequence is designed to evaluate the precision of the proposed method under large datasets. As it is concluded in the previous section that the chosen CNN model among the four tested will not influence the place recognition performance. Thus, one CNN model -NIN_ImageNet- is randomly chosen to extract the image features in this section.

As shown in Table 4.6, for all the sequences except day4, the place recognition results using inpainted images are better than the results using raw images, and are also better than the results of OLO. For sequence day4, OLO achieves 100% accuracy while the proposed method using sequence-inpainted images or using the raw images has a slightly worse performance -98.24% of accuracy- in terms of place recognition . Unexpectedly, the place recognition based on single-inpainted images of day4 gets the worst result among the three kinds of images based results, which is 0.88% lower than the result using raw images or sequence-inpainted images.

As for the comparison between the single-inpainted images based and sequence-inpainted images based place recognition, all the results using sequence-inpainted images are better than using the other kind of images except for sequence day3. For this sequence, the results of using inpainted images have a slight improvement -0.5%- than the sequence-inpainted image based result. As the sequence-inpainted images have such a good performance for place recognition, it is unexpected that the results based on sequence-inpainted images are always better than the results of DenseVLAD.

In summary, experiments on this extended dataset proves again the effectiveness of the dynamic objects removal methods for place recognition.

Table 4.7: The precision comparison of NIN_ImageNet based place recognition when using raw images and inpainted-images under three configurations (both the query and reference images inpainted, query images inpainted only and reference images inpainted only).

| Precision(%) | Single-inpainted | | | Sequence inpainted | | |
|---|---|---|---|---|---|---|
| | **both inpainted (query and reference)** | inpainted query/ raw reference | **inpainted reference/ raw query** | **both inpainted (query and reference)** | inpainted query/ raw reference | **inpainted reference/ raw query** |
| Dataset with the original query and reference sequences | | | | | | |
| day1 | **97.29** | 96.12 | **97.29** | **97.67** | 90.31 | **97.29** |
| day2 | **96.14** | 91.30 | **96.62** | **96.62** | 87.92 | **98.07** |
| day3 | **90.00** | 88.50 | **89.00** | **94.50** | 87.50 | **92.50** |
| day4 | **99.12** | 98.24 | **99.12** | **99.12** | 97.36 | **99.12** |
| day5 | **90.43** | 87.94 | **91.49** | **90.07** | 84.75 | **91.13** |
| Dataset with reversed query and reference sequences | | | | | | |
| day1 | **93.41** | 90.70 | **93.80** | **96.90** | 86.43 | **96.90** |
| day2 | **88.41** | 81.64 | **89.86** | **94.20** | 80.68 | **94.20** |
| day3 | **85.50** | 78.50 | **84.50** | **85.00** | 72.00 | **89.00** |
| day4 | **97.36** | 98.24 | **97.80** | **98.24** | 97.36 | **97.80** |
| day5 | **80.14** | 76.95 | **80.50** | **82.27** | 67.38 | **81.91** |

### 4.4.3/ PLACE RECOGNITION RESULTS CONSIDERING THE USE OF THE INPAINTED REFERENCE IMAGES ONLY TO IMPROVE TIME EFFICIENCY

To analyze how the dynamic objects removal step influences the performance of the place recognition depending on whether the inpainting step is applied on both the query and reference images or only on one of these two images, experiment in this part is setup as follows. Firstly, for each sequence, the place recognition results are computed separately using the three configurations shown in Fig. 4.17: 1) inpainted images for both images (i.e. query and reference), 2) inpainted query images and raw reference images and 3) raw query images and inpainted reference images. As two kinds of inpainted images can be generated depending on the inpainting method (single image or image sequence based), there will be six place recognition results by applying the different inpainted images for each sequence.

As for previous experiments, we resume the results into Table 4.7 based on NIN_ImageNet model under the extended ten sequences. The results in Table 4.7 show that the place recognition can have the best performance in most conditions by applying the approach on the inpainted reference images and raw query images. If the inpainted query images and raw reference images are used, the performance can be damaged whatever

the considered inpainting approach (single or sequence image based). Therefore, the proposed method can be improved by removing the dynamic objects only in the reference images resulting in saving time of image pre-processing.

## 4.5/ CONCLUSION AND FUTURE WORKS

This chapter proposed a new feature extraction system for visual place recognition. Before putting the whole images into a CNN model, they are submitted to semantic segmentation and image inpainting method to decrease the presence of dynamic information of images. This method improves the performance of place recognition in changing environments, especially in changing traffic conditions.

The robustness of features extracted by the proposed system relies on the performance of semantic segmentation and image inpainting. Image sequence inpainting contributes more than single image inpainting in our experiments, but both of them have some limitations as they can not perfectly inpaint the images and bring some artifacts. Even though, they can still improve the place recognition performance. Therefore, we have a deep research in Chapter 5 to see whether reducing instead of removing the noisy information of an image can improve the place recognition performance as well as be less time consuming.

# 5

# REDUCING IMAGE NOISE BASED VISUAL PLACE RECOGNITION

In Chapter 4, we showed the value of adding a step to detect and remove dynamic objects and background reconstruction ahead image retrieval for place recognition. The dynamic objects are recognized by semantic segmentation method firstly, and then the background information is restored by image inpainting method. However, the image inpainting method can not perfectly repair the background information and sometimes misreconstructed part of the dynamic objects. Even though, the place recognition performance has been improved by using this kinds of inpainted images.

Inspired by this, reducing instead of removing the image noise is proposed in this chapter. As image blurring is one of the simplest and most commonly method to reduce the image noise, Gaussian blurring is chosen to generate the blurred images. Based on the robust performance of removing the dynamic objects in last chapter, the so-called blurred-fcn image is generated by blurring the dynamic objects recognized by the semantic segmentation method - FCN. Besides, the whole blurred image is also generated by simply blurring the whole image. Experiments are done across all kinds of daytime conditions - day vs. day, night vs. night, and day vs. night, the results will show that the proposed method is an effective tool to tackle the place recognition across different conditions, especially in night vs. day conditions. As in previous Chapter 4, the results will be also compared with the state-of-the art DenseVLAD and OLO approaches.

## 5.1/ INTRODUCTION

Place recognition is still a challenging task when considering severe environmental changes. For daytime images, traffic conditions, weather and season changes will cause the drastically appearance change of the same place. For images under normal nighttime condition, the weak light of the street lamp instead of the natural light will cause the blur of the appearance of a place, even more if the images are captured under rainy night or night without lamps. If the appearance of a place under nighttime conditions is compared with its appearance under daytime condition, the difference is often too huge to make a correct recognition. Facing to the challenge, researches are proposed to decrease the difference of the appearance by generating synthesized images, such as changing the nighttime images to daytime ones [Anoosheh et al., 2019] [Porav et al., 2018], and removing the dynamic objects as in the work in Chapter 4, or improving the feature variations by combining the features during different daytime [Milford et al., 2012], while few works are proposed to weaken the features by simply blurring the images.

DeepDSAIR [Abolfazli Esfahani et al., 2019] is one of the works which proposes blurring the dynamic objects by training a new CNN model. This paper makes the dynamic objects blurred and enhances the static scene at the mean time by a Deblurred Semantic-Aware Image Representation (DSAIR) model. The first step of this model is to decide whether or not the raw image needs to be deblurred. If so, the trained deblurring model proposed in [Tao et al., 2018] is used to sharp the corners and edges of the raw image and gets a clearer image. Then, if the deblurred image should be processed by the semantic aware module, it would be sent to the Mask-RCNN model [He et al., 2017] and the dynamic objects of this image would be labelled by bounding boxes. Then, the output image with blurring dynamic information and enhanced background information is generated by the DSAIR model. This new image will be put in a deep CNN model to get a position and orientation output for achieving then the localization purpose.

Same as the work of DeepDSAIR, the proposed method in this chapter recognizes the dynamic objects by semantic segmentation and then blurred them, thus generating the blurred-fcn image with the blurry dynamic information. But this step is simpler in this proposed method: if uses FCN instead of the Mask-RCNN as the semantic segmentation method, the time efficiency is improved drastically while the precision of the recognition is

enough good. For the blurring part, the proposed method chooses the Gaussian blurring to tackle this task, which is simple and does not need to train a deep network. Moreover, the proposed method blurs the whole image which makes the whole image blurred, in contrary to the deblurring method to enhance the images, used in the DeepDSAIR method.

In the experiment part, these two kinds of blurred images - the blurred-fcn images or the whole blurred images are put as input of the place recognition system described in the section 3.1 (see Fig. 3.1). The experiments are divided into three parts: 1) place recognition under day vs. day condition; 2) place recognition under night vs. night condition; 3) place recognition under night vs. day condition. The three configurations almost cover all kinds of place recognition problem, and through the good performance, the proposed method proves its effectiveness in the condition that the reference images have better feature quality (clearer and sharper images) than the query images.

Inspired by the experiment of Chapter 4, the place recognition results are also compared by using both the blurred sequences with using either the query blurred or the reference blurred sequence. The results prove that blurring only the reference sequence is also an effective method for place recognition under most conditions.

The performance of the proposed method is also compared with the state-of-the-are works (DenseVLAD and OLO) by the precision parameter.

The key contributions of this work are:

- 1) Considering two kinds of blurred images: 1) the whole blurred-images that are blurred by the Gaussian blurring method in the whole images; 2) the blurred-fcn images that are generated by combining the semantic segmentation and image blurring method that blurs only the dynamic objects.
- 2) Proving that weaken the image feature can unexpectedly improve the place recognition performance.
- 3) Proposing a robust place recognition approach through blurring the reference sequence especially under night vs. day conditions.

The rest of this chapter is organized as follows. Firstly, the image blurring method and the proposed approach are introduced in Section 5.2. Section 5.3 presents the experi-

mental setup and exhibits the example images of the images generated by the proposed approach as well as the place recognition results. Finally, conclusions and future works close this chapter in section 5.4.

## 5.2/ The Proposed Approach

As considered in the previous chapter, deleting the noisy information such as dynamic objects can improve the place recognition. In this work, instead of completely removing the dynamic information, reducing the noisy information by image blurring method has been proposed. Two kinds of blurred images are generated. Firstly, simply blurring the whole image, which provides the whole blurred images. Then, blurring only the dynamic objects by combining the semantic segmentation method with the image blurring method, which provides the blurred-fcn images.

### 5.2.1/ Whole image blurring method

As image blurring can reduce the image noise, this part presents the effect of blurring a whole image and introduces some image blurring methods.

#### 5.2.1.1/ Image blurring

Image blurring is one of the simplest and most commonly used operation in image processing. One of its main purpose is to reduce image noise during image preprocessing, for example, removing some trivial details in the image before extracting large objects. The blurring of the image is usually realized by the convolution operation of the image. Image blurring is also called smoothing filtering.

Common image blurring methods can be listed as follows:

- 1) Normalized mean filter;
- 2) Bilateral filter;
- 3) Median filter;
- 4) Gaussian filter.

#### 5.2.1.2/ THE PRINCIPLE OF IMAGE BLURRING

**1) Normalized mean filter.** It is the simplest filter. The value of the output pixel is the mean value of the pixels in the corresponding kernel window (all pixels have equal weighting coefficients, though they can also be unequal). The core of the filter is as follows:

$$K = \frac{1}{K_{width} \cdot K_{height}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \cdot & \cdot & \cdot & \cdots & 1 \\ \cdot & \cdot & \cdot & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}. \tag{5.1}$$

Here, $K_{width}$ and $K_{height}$ represent the width and height of the kernel window, respectively.

**2) Median filter.** This filter is also relatively simple. The main idea of median filter is to run through an image pixel by pixel by a $n \times n$ template (or convolution kernel, mask), replacing each pixel with the median value of its neighbouring pixels in this template. For example, if using a $3 \times 3$ template, the calculation formula is as follows:

$$\begin{aligned} g(x,y) = median[&f(x-1,y-1), f(x-1,y+1), f(x+1,y-1), f(x,y), \\ &f(x+1,y+1), f(x-1,y), f(x,y-1), f(x+1,y), f(x,y+1)]. \end{aligned} \tag{5.2}$$

Here, f(x, y) is the gray value of the pixel (x, y) in the image, and g(x, y) is the value of the pixel after median filtering. Median filtering has a good suppression effect of salt and pepper noise.

**3) Bilateral filter.** It is good at smoothening images, reducing noise as well as preserving edges because it combines the characteristics of Gaussian Filter and Alpha-Trimmed Mean Filter, while considering the difference between both the spatial domain and range domain.

The spatial domain kernel $W_d(i, j, k, l)$ is formulated as follows:

$$\mathrm{w_d}(i,j,k,l) = exp\left(-\frac{(i-k)^2 + (j-l)^2}{2\sigma_d^2}\right), \tag{5.3}$$

Here, $(k, l)$ is the coordinate of the center point of the template $p(k, l)$ , and $(i, j)$ is the coordinate of another point of the template $q(i, j)$, and $\sigma_d$ is the standard deviation of the Gaussian function.

The range kernel is as follows:

$$w_r(i, j, k, l) = exp\left(-\frac{\|f(i, j) - f(k, l)\|^2}{2\sigma_r^2}\right). \tag{5.4}$$

Here, $f(i, j)$ is the pixel value of the point $q(i, j)$ and $f(k, l)$ is the pixel value of the point $p(k, l)$.

Therefore, the weight coefficient $w(i, j, k, l)$ of bilateral filter is defined as:

$$w(i, j, k, l) = w_d(i, j, k, l) \times w_r(i, j, k, l) = exp\left(-\frac{(i-k)^2 + (j-l)^2}{2\sigma_d^2} - \frac{\|f(i, j) - f(k, l)\|^2}{2\sigma_r^2}\right). \tag{5.5}$$

Then, the output pixel value $g(i, j)$ at point $q(i, j)$ of the bilateral filter is as follows:

$$g(i, j) = \frac{\sum_{k,l} f(k, l)w(i, j, k, l)}{\sum_{k,l} w(i, j, k, l)}. \tag{5.6}$$

Through a quantitative analysis, it can be found that when in a flat area, the difference between the pixel values of adjacent pixels is small, $w_r(i, j, k, l) \rightarrow 1$, which is equivalent to directly performing Gaussian blurring on this area. Therefore, the flat area is equivalent to Gaussian blur. And when the difference between the pixel values of neighboring pixels is large, $w_r(i, j, k, l) \rightarrow 0$, the Gaussian filtering effect will be greatly reduced, thus maintaining the details of the edge of the original image.

**4) Gaussian filter.** It is a linear smoothing filter which is suitable for eliminating Gaussian noise and is widely used in the noise reduction process of image processing. Generally speaking, Gaussian filtering is the weighted average process of the entire image. The value of each pixel is obtained by weighted average of itself and other pixel values in its neighborhood.

The idea of Gaussian filtering is to scan each pixel in the image with a template, and use the weighted average gray value of the pixels in the neighborhood determined by the template to replace the value of the center pixel of the template.

If a $3 \times 3$ template is used, the calculation formula is as follows:

$$\begin{aligned} g(x,y) = \{&f(x-1,y-1) + f(x-1,y+1) + f(x+1,y-1) + f(x+1,y+1) \\ &+[f(x-1,y) + f(x,y-1) + f(x+1,y) + f(x,y+1)] * 2 + f(x,y) * 4)\}/16; \end{aligned} \tag{5.7}$$

Here, f(x, y) is the gray value of the pixel (x, y) in the image, and g(x, y) is the value of the pixel after Gaussian filtering.

The output is the weighted average of the pixels in the template, i.e. the closer the distance of the surrounding point to the center point is, the greater the weight, and the farther the distance is, the smaller the weight. Therefore, compared to the mean filter, its smoothing effect is softer, and the edges are better preserved.

This chapter chooses the Gaussian filter to achieve the image blurring and the template size is set as $5 \times 5$. The effect of the image blurring can be seen in Fig. 5.4. And the whole image blurring method can be simply summarized as in Fig. 5.1.

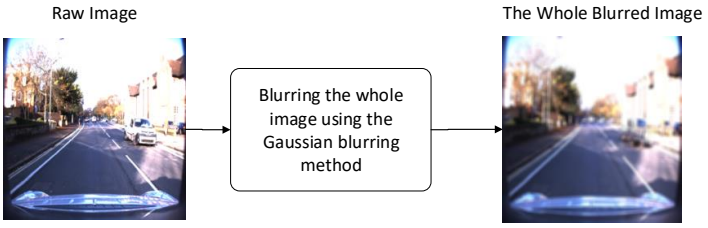



Figure 5.1: Illustrating the output of the whole blurred image. The raw image is blurred by the Gaussian blurring method to obtain the whole blurred image.

### 5.2.2/ DYNAMIC OBJECTS BLURRING METHOD

The results of the Chapter 4 have proved that the dynamic objects are noisy information for place recognition. To reduce it, dynamic objects blurring method is considered in this section. As illustrated in Fig. 5.2, the semantic segmentation method of FCN is used to recognize the dynamic objects, such as the car in the raw image and to label the dynamic objects as the colored parts in the mask. Then, the dynamic parts labelled in the mask are blurred by the image blurring method - Gaussian blurring. The output image with the blurred dynamic information is now called blurred-fcn image.

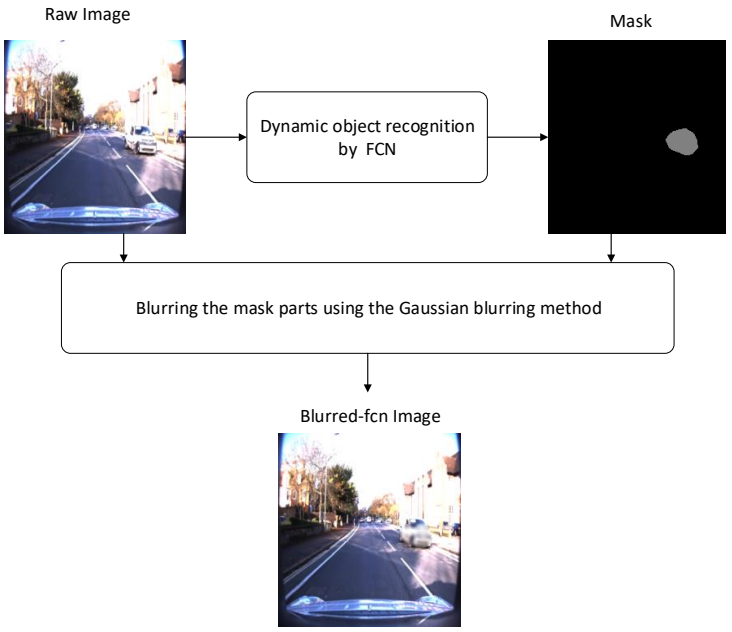



Figure 5.2: Illustration of the method to obtain the blurred-fcn image. The dynamic objects of the raw image are recognized by FCN and their corresponding masks are generated. Then Gaussian blurring method is applied on the raw image parts corresponding to the mask parts only to blur the dynamic objects.

### 5.2.3/ PLACE RECOGNITION SYSTEM

After generating the two kinds of blurred images, place recognition system as mentioned in Fig. 3.1 is used to recognize the place of the query image. In this chapter, NIN_ImageNet model is used to get the deep features of each input image. Then the nearest neighbouring searching is used to select the reference image that best matches to the query image.

In summary (see Fig. 5.3), the whole blurred images or the blurred-fcn images are generated by the proposed image blurring method (the yellow section) firstly, thus generating the whole blurred image sequence and the blurred-fcn image sequence. Then if the query and the reference images of the generated whole blurred image sequence are put into the place recognition system, the results based on the whole blurred images are obtained. If using the images in the generated blurred-fcn image sequence, the second results are obtained: place recognition results based on blurred-fcn images.

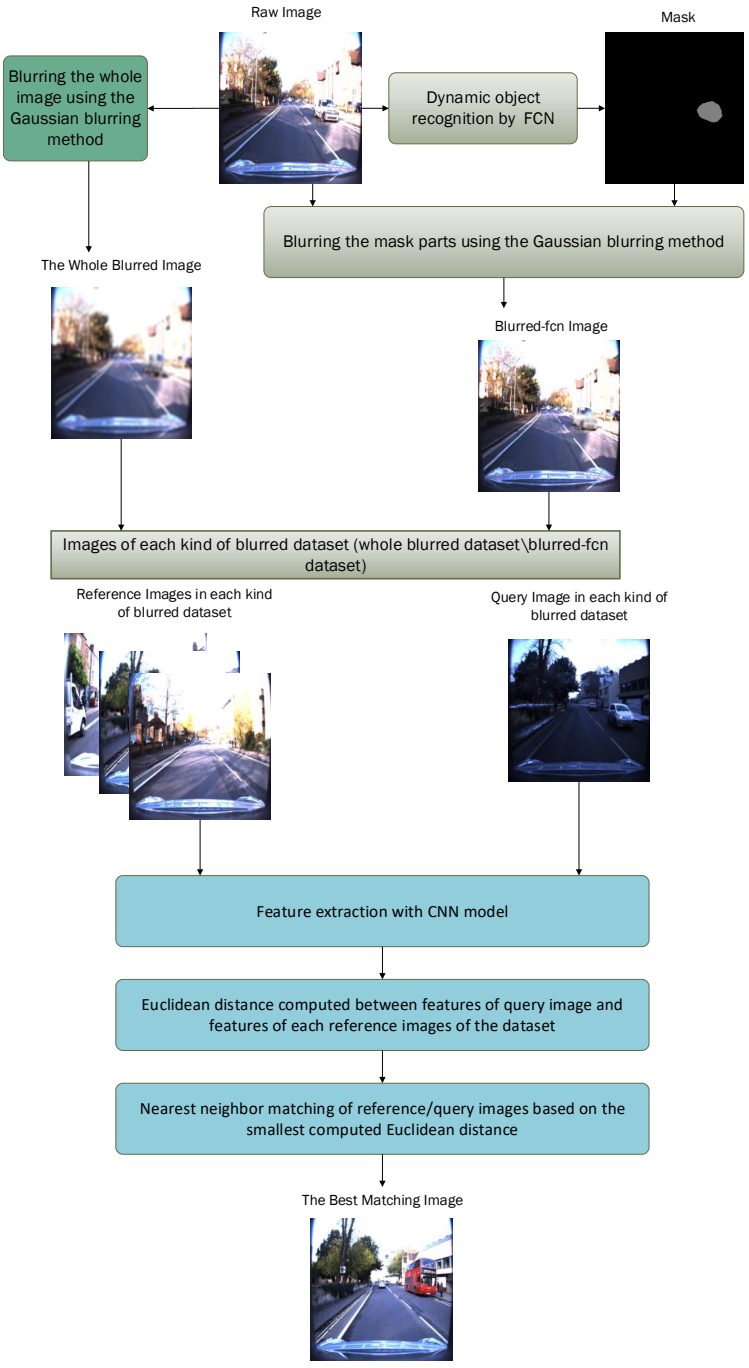



Figure 5.3: Summary of the proposed place recognition method. The best matching image comes from the result of place recognition step (blue section) using the query and reference images of the whole blurred image sequence or the generated blurred-fcn image sequence (generated by the image blurring step (yellow section)).

## 5.3/ EXPERIMENTAL SETUP AND RESULTS

This section shows the experimental setup and place recognition results obtained using the proposed method. The experiments are divided into three parts: 1) place recognition based on daytime images; 2) place recognition based on nighttime images; 3) place recognition based on night and day images. The datasets in each condition will be shown by exhibiting some examples of their raw images and the corresponding blurred images. As for Chapter 4, the results of the proposed methods are not only compared with themselves, but also compared with the advanced place recognition approaches (DenseVLAD and OLO).

### 5.3.1/ EXPERIMENTS BASED ON DAYTIME IMAGES

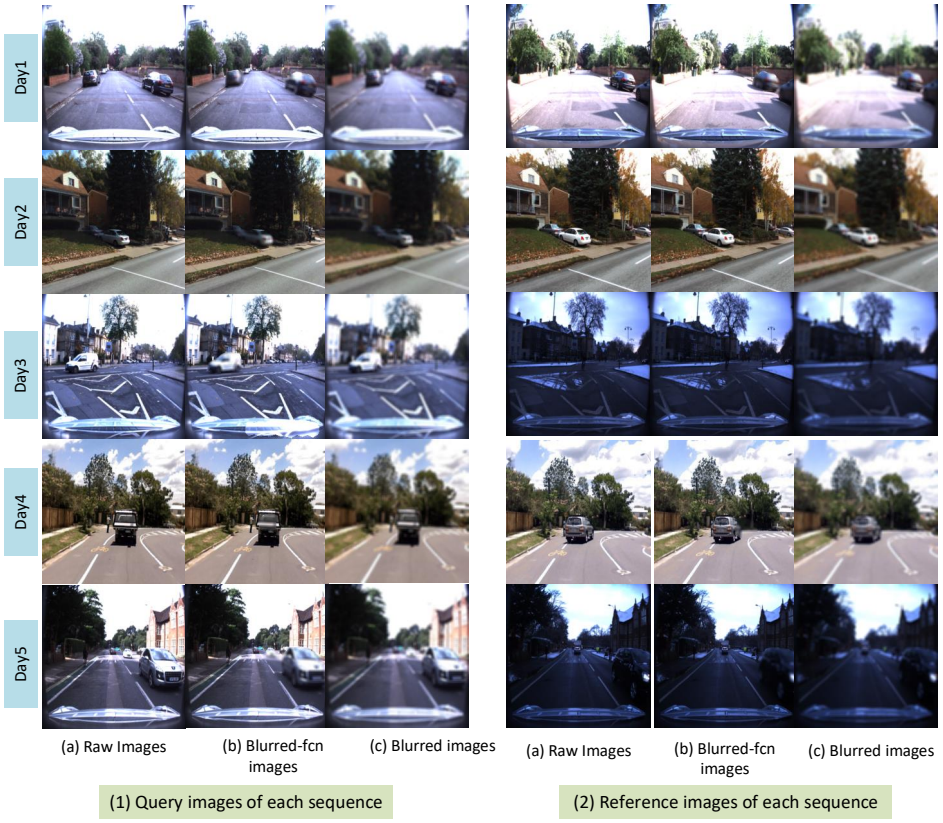



Figure 5.4: Examples of the query (1) and reference (2) images in the three kinds of generated image sequences: raw sequence (a), blurred-fcn sequence (b) and the whole blurred sequence (c).

Experiment based on daytime images includes two parts: 1) the place recognition by

using the whole blurred images, 2) the place recognition by using the blurred-fcn images. Results of the two parts are shown in the following.

As described in Table 3.1, five sequences are included in the daytime dataset which captured from the three available datasets: Oxford RobotCar, CMU and St.Lucia. Each sequence contains a query sub-sequence and a reference sub-sequence. 200 to 300 images, which represent the views in the same routes, were captured at different times and include in each sub-sequence. The images need to be numbered and preprocessed before using for experiment. More information of the dataset were introduced in section 3.2 of Chapter 3.

Examples of the query and reference images in the three kinds of image sequences: raw sequence, blurred-fcn sequence and the whole blurred sequence are shown in Fig. 5.4.

#### 5.3.1.1/ PLACE RECOGNITION RESULTS BASED ON THE WHOLE BLURRED IMAGES

Table 5.1: Precision comparison of NIN_ImageNet based place recognition when whole blurred-images are considered under three configurations (the query and reference images are blurred, only query images are blurred and only reference images are blurred) in the extended ten experimental sequences (i.e. original query and reference sequences and inversed query and reference sequences).

| Precision(%) | both blurred (query and reference) | query blurred /raw reference | reference blurred /raw query |
|---|---|---|---|
| Dataset with the original query and reference sequences | | | |
| day1 | **95.35** | 65.12 | **95.35** |
| day2 | 93.24 | 83.57 | **93.72** |
| day3 | **93.00** | 81.50 | 92.00 |
| day4 | **98.68** | 95.59 | **98.68** |
| day5 | **86.17** | 73.40 | 82.98 |
| Dataset with the inversed query and reference sequences | | | |
| day1 | 91.86 | 79.84 | **93.80** |
| day2 | 83.09 | 78.74 | **86.47** |
| day3 | 76.00 | 46.00 | **85.00** |
| day4 | 97.36 | 96.04 | **97.80** |
| day5 | 73.76 | 39.36 | **78.72** |

As it has been pointed out in Chapter 4, applying the preprocessing (here blurring) on both query and reference images or only on one of these images may have an influence on the performance of place recognition.

Table 5.1 shows the place recognition results considering 3 configurations: 1) query and reference blurred images are used, 2) query blurred and raw reference images are used, and 3) raw query and blurred reference images are used on each of the extended ten daytime datasets (as illustrated in Fig. 4.16, the dataset are extended by inversing the query and and reference images in each sequence). The best place recognition results based on the whole blurred images are obtained in most conditions (except the sequences of original day3 and day5) by applying the raw query and blurred reference images. For the sequences of original day3 and day5, even the results considering the third configuration are worse than those considering the first configuration (1.00% and 3.19% performance degradation respectively), it is worth to applying this kind of configuration (raw query and blurred reference images) for obtaining time efficiency.

#### 5.3.1.2/ PLACE RECOGNITION RESULTS BASED ON DYNAMIC OBJECTS BLURRED IMAGES

Table 5.2: Precision comparison of NIN_ImageNet based place recognition when blurred-fcn images are considered under three configurations (the query and reference images are both blurred, only query images are blurred and only reference images are blurred) in the extended ten experimental sequences (i.e. original query and reference sequences and inversed query and reference sequences).

| Precision(%) | both blurred (query and reference) | query blurred /raw reference | reference blurred /raw query |
|---|---|---|---|
| Dataset with the original query and reference sequences | | | |
| day1 | 96.51 | 96.51 | 96.51 |
| day2 | 92.75 | 92.75 | 92.75 |
| day3 | 89.50 | 88.50 | **90.00** |
| day4 | 97.36 | **98.24** | 96.92 |
| day5 | **88.30** | 87.59 | 87.59 |
| Dataset with the inversed query and reference sequences | | | |
| day1 | **92.64** | 90.70 | 92.25 |
| day2 | 82.13 | 82.13 | 82.13 |
| day3 | **83.00** | 78.00 | 82.00 |
| day4 | 97.80 | **98.24** | **98.24** |
| day5 | **79.43** | 76.95 | **79.43** |

It can be seen from the example images of the Fig. 5.4 that the images processed by the dynamic objects blurring method have only blurred the vehicles (and pedestrians) and keep the remaining information unchanged. The results presented in Table 5.2 do not

show any regular pattern. Using both query and reference blurred-fcn sequences or either one of the blurred-fcn sequence has no influence on place recognition for the original sequences day1, day2 and reversed day2. The best performance is obtained using both the query and reference blurred-fcn sequence for the original day5 and reversed day1, day3 and day5 sequences. The best performance is obtained using only the blurred-fcn query sequence for the original and reversed day4 sequences. The best performance is obtained using only the blurred-fcn reference sequence for the original day3, reversed day4 and day5 sequences. Through these results, it can be concluded that using only the reference blurred-fcn query sequence can not always improve the place recognition performance. Using only the blurred-fcn query sequence, only the blurred-fcn reference sequence or both the blurred-fcn query and reference sequence has no impact on place recognition. It is perhaps the quantity and the variations of the dynamic objects between the query and reference images that influence the results of place recognition based on the blurred-fcn images.

#### 5.3.1.3/ COMPARISON OF THE PROPOSED METHOD WITH ADVANCED METHODS OF THE LITERATURE

Table 5.3: Precision comparison of NIN_ImageNet based place recognition using the raw images, the blurred-fcn images and the whole blurred-images (blurring only the reference sequence) with three advanced methods in the extended ten experimental sequences (i.e. original query and reference sequences and inversed query and reference sequences) .

| Precision(%) | ours | | | other methods | | |
|---|---|---|---|---|---|---|
| | raw | reference blurred-fcn | reference blurred | sequence-inpainted (reference) | DenseVLAD | OLO |
| Dataset with the original query and reference sequences | | | | | | |
| day1 | 96.90 | 96.51 | 95.35 | **97.29** | 89.15 | 73.64 |
| day2 | 92.75 | 92.75 | **93.72** | **98.07** | 93.24 | 93.24 |
| day3 | 89.00 | **90.00** | **92.00** | **92.50** | 73.00 | 81.50 |
| day4 | 97.80 | 96.92 | **98.68** | **99.12** | 96.92 | 100.00 |
| day5 | 86.88 | **87.59** | 82.98 | **91.13** | 38.30 | 66.67 |
| Dataset with reversed query and reference sequences | | | | | | |
| day1 | 90.70 | **92.25** | **93.80** | **96.90** | 96.51 | 83.33 |
| day2 | 82.61 | 82.13 | **86.47** | **94.20** | 93.72 | 86.47 |
| day3 | 78.50 | **82.00** | **85.00** | **89.00** | 66.50 | 79.00 |
| day4 | 98.24 | 98.24 | 97.80 | 97.80 | 99.12 | 100.00 |
| day5 | 77.66 | **79.43** | **78.72** | **81.91** | 34.40 | 67.38 |

Even if using the blurred-fcn reference sequence which can not guarantee the best per-

formance of place recognition, the place recognition method based on the blurred-fcn reference sequence still has its advantage and is a good compromise considering processing time. At the same time, the place recognition method based on the whole blurred reference sequence can not only improve the time efficiency but also improve the place recognition performance at the most time. Therefore, the results of the place recognition using only the reference blurrred-fcn and the whole blurred sequences are compared with the results based on the raw images and two other advanced methods (DenseVLAD and OLO) in this part.

Even if the results using the blurred images are better than using the raw images in most conditions, it is hard to define these conditions. Conclusions cannot be made if the blurred images based results are analyzed separately, but connecting them with the raw images based results makes it easier. Regarding the results using the raw images, even using the same sequence, but just reversing the query and reference sequences, the place recognition performance can have a big difference. For example, the results using the original sequences day2, day3 and day5 are obvious better than the results based on the inversed sequences. From the example images of the Fig. 5.4 (2), the original reference images of day3 and day5 are in the winter conditions and the features are less distinguishable. Then using this winter sequence as the query sequence, the results are not very good (as shown in the reversed day3 and day5 results in Table 5.3); using this more distinguishable summer sequence as the query sequence, the results are significantly improved (as shown in the original day3 and day5 results in Table 5.3). Then it can be concluded that the distinguishability of the query image determines the place recognition performance.

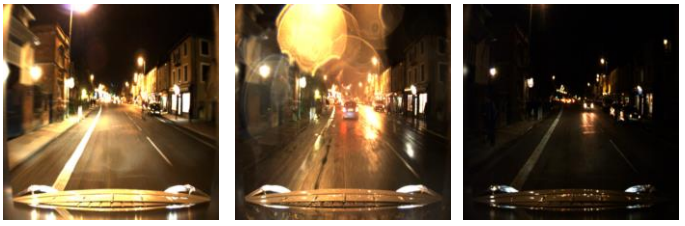
Then this conclusion can be connected with the place recognition performance based on the reference blurred images. From the results based on the original day4, the reversed sequences day1, day2, day3, day5 of Table 5.3, it can be seen that when using the sequence whose features are less distinguishable as the query sequence, blurring the reference sequences can improve the place recognition performance. To evaluate this conclusion, more results are shown in the following parts.

Though in these conditions blurring the whole image of the reference sequence can be useful, it is less effective compared to the dynamic objects removing and inpainting method proposed in Chapter 4. Table 5.3 shows only the results obtained using the se-

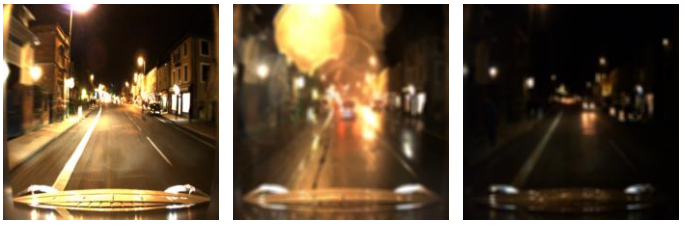
quence image inpainting method, because the results of applying the sequence-inpainted images can always performs better than applying the single-inpainted images. The same as the proposed blurring method, the dynamic objects inpainting method is also applied only to the reference sequence. The blurring method performs not as good as the proposed method in Chapter 4, but is much more simple and less time consuming. And comparing with the two other place recognition approaches (DenseVLAD and OLO), the proposed approach using only the whole blurred reference images has a better performance for most sequences.

As for the method of blurring the dynamic objects, the results in Table 5.3 can not show its competitiveness. But the sequences day3 and day5 are under much more busy traffic conditions compared to the other sequences (as it can be seen in Fig. 5.4), the results using the blurred-fcn images are better than using the raw images and even better than using the blurred images (in both the original and reversed day5 sequences). It can perhaps prove the effectiveness of this method in this busy traffic condition.

### 5.3.2/ EXPERIMENTS BASED ON NIGHTTIME IMAGES



(a) The raw night images under three conditons (normal night, rainy night, very dark night form left to right).



(b) The corresponding blurred images of the three kinds of night images.

Figure 5.5: Examples of the whole blurred images in the three nighttime sequences and their corresponding raw night sequences.

Nighttime images are always under low-quality conditions, they are chosen to evaluate

Table 5.4: Precision comparison of NIN_ImageNet based place recognition using the raw images, the whole blurred-images under three configurations (both the query and reference images are blurred, only query images are blurred and only reference images are blurred) with the advanced method in the extended six nighttime datasets.

| Precision(%) | raw | blurred image | | | advanced method |
|---|---|---|---|---|---|
| | | both blurred | query blurred | reference blurred | DenseVLAD |
| night-rain vs. night | 63.56 | **69.92** | 35.59 | **76.69** | 37.29 |
| night-darker vs. night | 50.85 | 50.85 | 20.34 | **69.92** | 88.98 |
| night-darker vs. night-rain | 23.73 | **33.47** | 16.53 | **41.53** | 32.20 |
| Dataset with the reversed query and reference sequences | | | | | |
| night vs. night-rain | 54.24 | **55.08** | 54.24 | 49.58 | 35.59 |
| night vs. night-darker | 88.98 | **89.83** | 88.14 | 83.05 | 87.71 |
| night-rain vs. night-darker | 76.69 | **81.36** | 52.12 | **78.39** | 30.93 |

if blurring the images under low-quality conditions can still improve the place recognition performance. Fig. 5.5 shows the nighttime dataset under three conditions (normal night, rainy night, and night with little luminance) used in the experiment (details about the nighttime dataset can be found in Table 3.2). Some examples of the raw images from the three datasets are shown in top line (a). And their corresponding whole blurred images are shown in bottom line (b). The blurred-fcn images are not generated in this experiment as the dynamic information are not obvious in the nighttime conditions.

The experiments are divided into three parts in Table 5.4. Firstly, results obtained using both blurred query and reference sequences, and those either of the blurred query or reference sequence are compared. From the results of Table 5.4, it can be seen that the place recognition performance is stable and robust if using both the blurred query and reference sequences. But if considering the time efficiency, using only the reference blurred sequence is a better choice especially if the feature of reference sequence has better quality. Then, the results using the proposed blurred images are compared with the results using the raw images: the best results obtained with the blurred images are always better than those obtained with the raw images. It can conclude that the proposed image blurring method has better accuracy for place recognition.

Finally, by comparing the results obtained by the proposed approach with the state-of-the-art place recognition approach, DenseVLAD, it can be seen that DenseVLAD has its advantage in some condition (low-quality query images vs. good quality reference images), but the proposed method has an obvious improvement in most conditions.

### 5.3.3/ EXPERIMENTS BASED ON NIGHT AND DAY IMAGES

(a) The night images under three conditons as the query image.

(b) The blurred-fcn image (middle) or the whole blurred image (right) of the same daytime sequence can be used as the reference images.

Figure 5.6: Illustration of dataset setup for day and night place recognition.

Table 5.5: Precision comparison of NIN_ImageNet based place recognition using the raw images and the whole blurred-images with the advanced method in the night and day dataset.

| Precision(%) | ours | | advanced method |
|---|---|---|---|
| | raw | day blurred | DenseVLAD |
| night vs. day | 51.69 | 59.75 | 19.92 |
| night-rain vs. day | 22.88 | 23.73 | 10.88 |
| night-darker vs. day | 5.51 | 10.59 | 16.32 |

In the previous experiment, the proposed method has proved its accuracy if the reference image has better quality than the query image. As the quantitative analysis of the image quality has not been proposed in this chapter, the aim of the night vs. day place recognition experiment is to give an intuitive sense to the readers. As it can be seen from the

Fig. 5.6, the query sequences are chosen from the three nighttime datasets described in the previous part, and the reference sequence is chosen from one daytime sequence which represents the same places corresponding to the images in the query nighttime sequence. And comparing the nighttime and daytime images, it can be clearly shown that the feature of day image is better than the feature of nighttime image. Therefore, this experiment is under the configuration that the reference (daytime) sequence has better quality than the query (nighttime) sequence.

Fig. 5.6 (b) shows the effect of applying the whole image blurring method (the whole blurred images in the right) and the dynamic objects blurring method (the blurred-fcn images in middle) in the raw daytime images (left). As the daytime sequence has little dynamic objects, which have little impact on the place recognition performance, only the whole blurred images is applied in the experiment.

Table 5.5 shows the results using the raw images, those using blurred reference but raw query images and the results of DenseVLAD also. It verifies that blurring the reference images and keeping the raw query images can improve the place recognition performance in the condition that reference sequence has better quality than the query sequence. DenseVLAD is good at recognizing features when the query image has a little distinguished features and the reference images are captured in a pretty good condition. Therefore, in the case of place recognition under night-darker vs. day condition, DenseVLAD has better performance. But in other two conditions, the proposed method shows its accuracy by dramatically improving the results compared to DenseVLAD.

## 5.4/ Conclusion and Future Works

In this chapter, we proposed a method that integrates an image noise reducing step to improve the place recognition performance. By using image blurring method to reduce the noise information of the whole image, the whole blurred images are generated. In case of better quality of reference images with respect to query images (as night vs. day condition), processing the reference images to form the whole blurred images and keep the query images unchanged, the place recognition was improved obviously. In other conditions, blurring both the reference and query images can improve the place recognition performance at most times. By combining the image blurring method with the seman-

tic segmentation method, the blurred-fcn images with blurry dynamic information (cars, tracks, and pedestrians) can be obtained. Only in the busy traffic conditions, applying the blurred-fcn can improve the place recognition performance.

However, the quality of the images can not be described quantitatively now. To overcome this limitation, we plan to have a deep research on how to define the quality of image feature in the future work to analyse more deeply the compared results.

# 6

# Day and Night Place Recognition Based on Low-quality Images

As it has been pointed out in section 5.3.1.3 of Chapter 5, the quality of the query image influences the proposed place recognition method. As the quality of day and night images used in this chapter have a large difference, the day and night place recognition can be divided into two problems: 1) night vs. day place recognition and 2) day vs. night place recognition. The first problem deals with selecting the reference image from the daytime images sequence which represents the same place as the nighttime query image. In the contrary, the query image and reference sequence in the second problem are the daytime and nighttime images, respectively.

The feature similarity achievement approach is used to tackle night vs. day place recognition, which is detailed in Fig. 6.1 (a). The principle is to balance the quality of the nighttime and daytime images features. This is achieved by: 1) decreasing the quality of the daytime images by image blurring method (as done in Chapter 5); 2) increasing the quality of the nighttime images by an image enhancement method. The dynamic objects removal method presented in Chapter 4 is then combined with the similarity achievement approach to evaluate the influence of the dynamic objects on night vs. day place recognition.

And the problem of day vs. night place recognition is addressed using the image enhancement method and the image inpainting method (as described in Chapter 4) to improve the quality of the nighttime images, as detailed in Fig. 6.1 (b).

The approaches to tackle the two problems are tested on the Oxford RobotCar dataset,

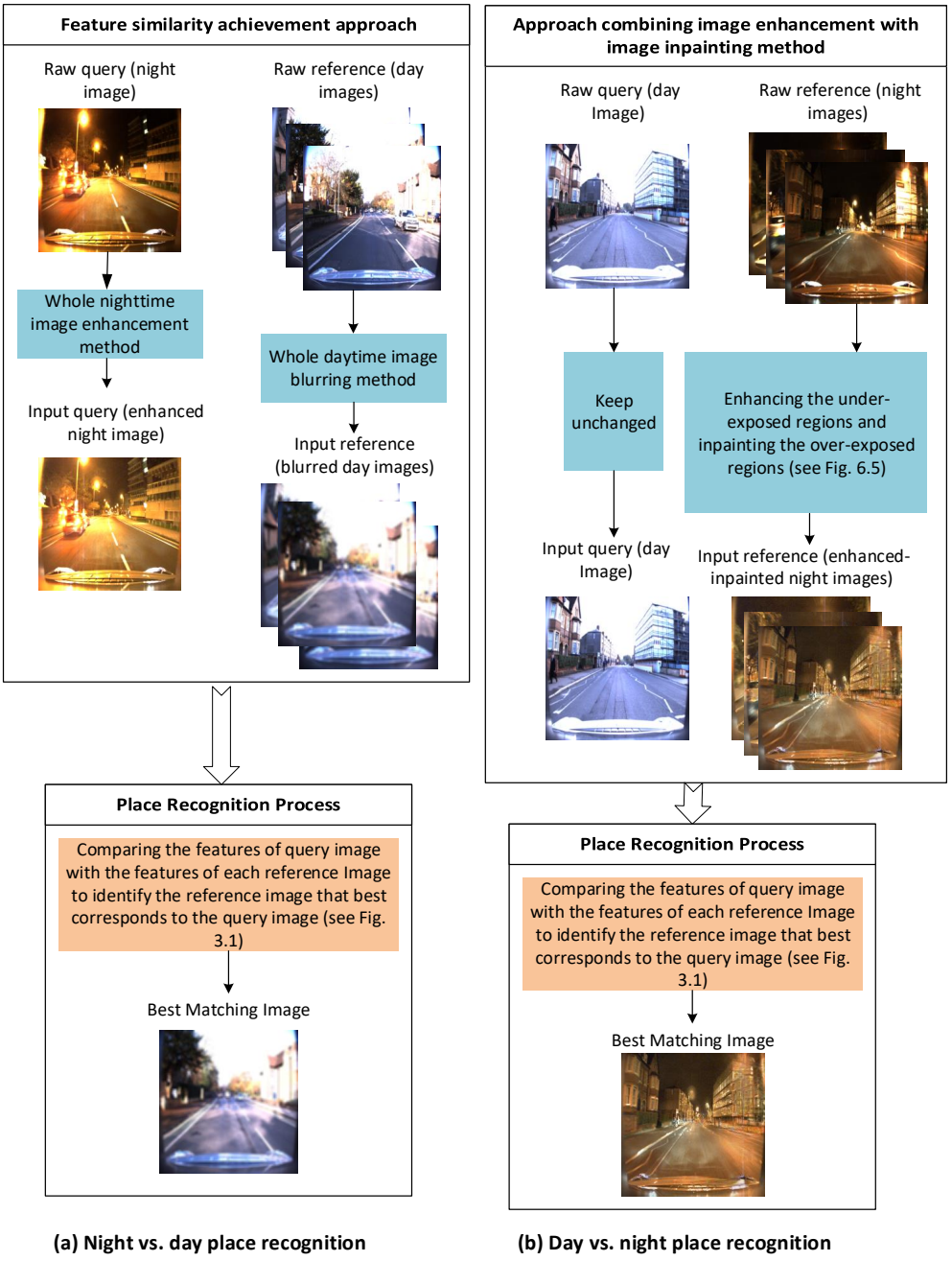



Figure 6.1: The day and night place recognition is divided into two problems in this chapter: (a) night vs. day place recognition, (b) day v.s night place recognition. The feature similarity achievement approach is used to tackle problem (a). Image enhancement and image inpainting method are used to tackle problem (b).

where three low-quality night sequences that represent the same routes are selected as query sequences among 8 night sequences, and their corresponding day sequence is selected as a reference sequence for problem 1. The results obtained with the proposed ap-

proaches are better than those obtained with the raw nighttime images. The results of our proposed place recognition system are also compared with two state-of-art place recognition methods: ToDayGAN [Anoosheh et al., 2019] and DenseVLAD [Torii et al., 2015].

## 6.1/ INTRODUCTION

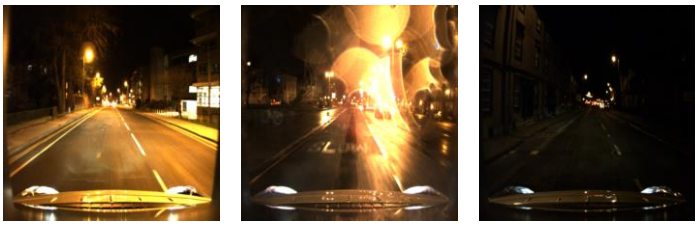
Ideally, a robust visual place recognition method should recognize the place no matter the environmental conditions. This process faces a lot of real difficulties, such as variations in illumination, time of day, weather, season and camera position, etc. Image retrieval is often used to achieve this process: finding the most similar reference image compared to a query image. In this case, it is difficult to find a correct matching when dramatically appearance changes between reference images and the query one occur. Facing this big challenge, tremendous progress has been made in visual place recognition [Anoosheh et al., 2019] [Arandjelovic et al., 2016] [Cummins et al., 2008] [Milford et al., 2012] [Porav et al., 2018] [Sattler et al., 2017] [Svärm et al., 2017] [Torii et al., 2015] in recent years.

Enlightened by the fact that place recognition is much more easy when the query and reference images have similar appearance, many kinds of research [Anoosheh et al., 2019] [Porav et al., 2018] [Torii et al., 2015] focus on translating the reference image in terms of viewpoint or appearance into a synthesized image which is similar to the query image.

The place recognition approach -DenseVLAD- in [Torii et al., 2015] aims at generating off-line synthesis views of reference images that have the same view as the query image. The improved results prove the success of considering same viewpoint strategy in place recognition.

With the development of GANs (generative adversarial networks), the robustness of GAN-based image-to-image translation methods [Anoosheh et al., 2018] [Choi et al., 2018] [Ignatov et al., 2018] [Liu et al., 2017] [Zhu et al., 2017] make it possible to generate drastically appearance changing synthesized images, e.g transform day to night, summer to winter, etc. Works in [Porav et al., 2018] and ToDayGAN [Anoosheh et al., 2019] transfer nighttime images to daytime ones to deal with day-night changes in visual place recognition.

ToDayGAN, the day and night place recognition method that generates the synthesized daytime image, is as follows. Firstly, it trains an image translation model targeting in day and night image domains and translates the input query images from night to day. Then the features of both the translated images and reference images are generated by the method of DenseVLAD. Through the nearest neighboring search, a closest matching reference image of each query image is obtained, and the day and night localization problem is solved. However, the quality of synthesized images depends on the quality of dataset images and not all the dataset images have enough good quality. When using low-quality night-time images, ToDayGAN has poor performance in generating daytime synthesized images using its pre-trained networks.



(a) Night image (b) Night-rain image (c) Night image under little illumination

Figure 6.2: Low-quality night-time images used in this chapter.

Since low-quality images can not be avoided in real visual place recognition tasks, this chapter aims at changing the quality of image features and proposes a feature selection process.

For night vs. day place recognition, a novel feature similarity approach instead of the appearance similarity (like transforming night image to day) between the query and reference images is proposed. To achieve this objective, we observed that daytime images in the dataset usually have good quality and night-time images have a worse quality due to the lower illumination like images shown in Fig. 6.2 (a). The night-time images can even have terrible quality when they are obtained in severe weather conditions, like images shown in Fig. 6.2 (b) (c). Since the quality of daytime and night-time images have a huge difference, our method aims to achieve the feature similarity into two ways: 1) improving the quality of night-time images by image enhancement method, and 2) ”damaging” the quality of daytime images by image blurring method. In this way, the daytime and night-

time images will not show any similarity in the appearance. But through the improved place recognition performance, we can suppose that they have favor feature similarity.

To evaluate the performance of our proposed method, the night vs. day place recognition results based on the preprocessed images are not only compared with the results based on the raw images but also compared with the results of two state-of-art approaches: DenseVLAD and ToDayGAN.

After proving the performance of the proposed method, further research is proposed in this chapter. As the negative influence of dynamic objects, as cars and pedestrians, on the daytime place recognition was proved in Chapter 4, this work continues the research and shows the influence of the dynamic objects in the images on the day and night place recognition through the dynamic objects removing method.

As for the day vs. night place recognition, the feature selection process is modified in order to enhance the nighttime images as follows: 1) increasing the useful information of night-time images by enhancing their under-exposed regions, 2) repairing the hard light regions by image inpainting. Three kinds of enhanced nighttime images are generated by the proposed method, as shown in Fig. 6.5 (and detailed in section 6.3.2). This process is only used for nighttime images, as Chapter 5 has concluded that if the query image has a good quality, there is no need to damage it.

The experiments part is similar as in the night vs. day place recognition. The results obtained with the three enhanced images are compared with those obtained with raw images firstly. Then, the best result is compared to those provided by the two state-of-art place recognition methods: ToDayGAN and DenseVLAD.

The key contributions of this work are:

- 1) Dividing the place recognition problem into two parts: night vs. day and day vs. night place recognition.
- 2) Proposing a novel approach to achieve the feature similarity of the query and the reference images for night vs. day place recognition. And proposing different image-enhancing methods to effectively improve the robustness of low-quality nighttime image features for day vs. night place recognition.
- 3) Explaining the influence of the dynamic objects on day and night place recogni-

tion.

The rest of this chapter is organized as follows. Firstly, the used image enhancement method is introduced in Section 6.2. Then, in Section 6.3, the proposed approaches for night vs. day and day vs. night place recognition problems are explained in detail. Section 6.4 exhibits examples of the images generated by the proposed approaches and presents and analyzes the obtained results. Finally, conclusions and future works close this chapter in section 6.5.

## 6.2/ PRE-PROCESSING USED METHODS

To solve the day and night place recognition problem, image enhancement method, image inpainting method and image blurring method are introduced. As image inpainting and image blurring have already been described in the previous Chapters, this section will briefly describes the image enhancement method.

### 6.2.1/ THE USED NIGHTTIME IMAGE ENHANCEMENT METHOD

A low-light image enhancement method designed in work [Ying et al., 2017] is chosen to enhance the image. The framework of this algorithm is based on a dual-exposure fusion method. The first step generates one appropriate exposure image. The second step fuses the generated image and input image into the enhanced result which is defined as

$$\mathbf{R}^c = \hat{\mathbf{W}} \circ \mathbf{P}^c + (1 - \hat{\mathbf{W}}) \circ g\left(\mathbf{P}^c, \hat{k}\right). \tag{6.1}$$

$\mathbf{P}$ is the input image, $\mathbf{R}$ is the enhanced result and $c$ is the index of three color channels. The enhancement algorithm consists of three parts: dual-exposure generator ($g$), dual-exposure sampler ($\hat{k}$) and dual-exposure evaluator ($\hat{\mathbf{W}}$). In the following subsections, the three parts will be introduced one by one.

#### 6.2.1.1/ DUAL-EXPOSURE GENERATOR

The dual-exposure generator aims to transform the input image into a new one with different exposure. Then, by combining this generated image with the input image, the

enhanced image can be obtained. Given the exposure ratio ($k$) and the dual-exposure generator ($g$), the image ($\mathbf{P}_1$) can be generated according to the input image ($\mathbf{P}$).

$$\mathbf{P}_1 = g(\mathbf{P}, k) = \beta \mathbf{P}_0^{\gamma}, \tag{6.2}$$

where $\beta$ and $\gamma$ are parameters determined by the camera characteristics and the given exposure ratio $k$, which can be computed by solving the following function:

$$f(kE) = \beta f(E)^{\gamma}. \tag{6.3}$$

$f$ is defined by:

$$f(E) = \begin{cases} e^{b(1-E^a)}, & \text{if } \gamma \neq 1, \\ E^c, & \text{if } \gamma = 1. \end{cases} \tag{6.4}$$

Since $\gamma \neq 1$ is the case that is mainly considered, this function can be solved as:

$$a = \log_k \gamma, \quad b = \frac{\ln \beta}{1-\gamma}. \tag{6.5}$$

Here, $a$ and $b$ are two parameters that are set in most cameras as $a = -0.3293, b = 1.1258$.

Therefore, $\beta$ and $\gamma$ can be obtained through $a$ and $b$, thus obtaining $g$:

$$g(\mathbf{P}, k) = e^{b(1-k^a)} \mathbf{P}^{(k^a)}, \tag{6.6}$$

Then, given an exposure ratio (k), the generated image ($\mathbf{P}_1$) can be also calculated.

#### 6.2.1.2/ DUAL-EXPOSURE SAMPLER

The image enhancement method in [Ying et al., 2017] only generates an image, so the optimal exposure ratio $k$ of this generated image needs to be able to recover as many as possible information in the under-exposed regions of the original image. The aim of the dual-exposure sampler is to find the appropriate exposure ratio $k$ by solving a one-dimensional problem. The following part details this process.

Firstly, an image ($\mathbf{Q}$) with only under-exposed pixels of the input image ($\mathbf{P}$) can be obtained by the equation:

$$\mathbf{Q} = \{\mathbf{P}(x) | \mathbf{T}(x) < 0.5\}. \tag{6.7}$$

Here, $\mathbf{P}(x)$ is the pixel value in the position of point $x$ in the input image $\mathbf{P}$ and $\mathbf{T}(x)$ is the pixel value in the position of point $x$ in map $\mathbf{T}$. The method to estimate the illumination map $\mathbf{T}$ will be introduced in section 6.2.1.3. When estimating $k$, the brightness component $\mathbf{B}$, instead of the color information, needs to be considered, so the geometric mean of red, green and blue channels of $\mathbf{Q}$ can be used to define $\mathbf{B}$:

$$\mathbf{B} = \sqrt[3]{\mathbf{Q}_r \circ \mathbf{Q}_g \circ \mathbf{Q}_b}. \tag{6.8}$$

Here, $\circ$ is the composition function which is defined by $(\mathbf{Q}_r \circ \mathbf{Q}_g)(x) = \mathbf{Q}_r(\mathbf{Q}_g(x))$. Then, the image entropy can be defined as:

$$\mathcal{H}(\mathbf{B}) = -\sum_{i=1}^{N} p_i \cdot \log_2 p_i. \tag{6.9}$$

$N$ is the number of bins which is always fixed to $256$ and $p_i$ represents the number of data of $\mathbf{B}$ valued in $\left[\frac{i}{N}, \frac{i+1}{N}\right)$.

Since the well-exposed images contain more information and their image entropy is higher than that of the under-exposed images, maximizing the image entropy of the enhancement brightness can obtain the optimal exposure ratio $\hat{k}$:

$$\hat{k} = \arg\max_k \mathcal{H}(g(\mathbf{B}, k)). \tag{6.10}$$

When increasing $k$, image entropy increases first and then decreases, so $\hat{k}$ is a one-dimensional minimizing problem. Then, by changing the value of $k$, the appropriate $k$ can be obtained finally.

#### 6.2.1.3/ DUAL-EXPOSURE EVALUATOR

When fusing the input image ($\mathbf{P}$) and the generated image ($\mathbf{P}_1$), the well-exposed regions of the input image can be preserved while the under-exposed regions of the input image should be replaced by the same regions in the generated image. So it is important to design the weight matrix $\hat{\mathbf{W}}$ to choose the regions.

As the weight matrix has positive correlation with the scene illumination, it is defined as:

$$\hat{\mathbf{W}} = \mathbf{T}^{\mu} \tag{6.11}$$

where $\mathbf{T}$ is the scene illumination map and $\mu$ controls the enhancement degree. When $\mu = 0$, the output image has no enhancement and is equal to the input image. When $\mu = 1$, the whole input image is enhanced including the under-exposed regions and well-exposed regions. When $\mu > 1$, the well-exposed regions may be over-exposed and the enhanced result image may lose details. As in [Ying et al., 2017], $\mu$ is set to $0.5$. The estimation of scene illumination map $\mathbf{T}$ can be seen as an optimization problem.

a. Optimization Problem: The scene illumination uses lightness component for estimation. And initial illumination map

$$\mathbf{L}(x) = \max_{c \in \{R,G,B\}} \mathbf{P}_c(x) \tag{6.12}$$

is adopted as lightness component. In this equation, $x$ is each individual pixel of the scene. Therefore, the following optimization formulation can be used to refine $\mathbf{T}$.

$$\min_{\mathbf{T}} ||\mathbf{T} - \mathbf{L}||_2^2 + \lambda ||\mathbf{M} \circ \nabla \mathbf{T}||_1, \tag{6.13}$$

where $\lambda$ is a coefficient and $\mathbf{M}$ is a weight matrix. The first term of this equation is to make the initial map $\mathbf{L}$ and the scene illumination map $\mathbf{T}$ infinite approximated, and the second term controls the smoothness of $\mathbf{T}$. The design of the weight matrix $\mathbf{M}$ is the key for the $\mathbf{T}$ refinement and it is designed as:

$$\mathbf{M}_d(x) = \frac{1}{|\sum_{y \in w(x)} \nabla_d \mathbf{L}(y)| + \epsilon}, \tag{6.14}$$

where $w(x)$ is the local window surrounding the center pixel x. To avoid the zero denominator, setting a very small constant $\epsilon$ is necessary.

b. Closed-Form Solution: To solve the optimization problem as in [Guo et al., 2017], Eq. 6.13 is approximated to:

$$\min_{\mathbf{T}} \sum_x \left( (\mathbf{T}(x) - \mathbf{L}(x))^2 + \lambda \sum_{d \in \{h,v\}} \frac{\mathbf{M}_d(x) \, (\nabla_d \mathbf{T}(x))^2}{|\nabla_d \mathbf{L}(x)| + \epsilon} \right). \tag{6.15}$$

Using $\mathbf{m}_d$, $\mathbf{l}$, $\mathbf{t}$ and $\nabla_d\mathbf{l}$ representing $\mathbf{M}_d$, $\mathbf{L}$, $\mathbf{T}$ and $\nabla_d\mathbf{L}$ respectively, the problem is translated to a linear function solving problem:

$$\left(\mathbf{I} + \lambda \sum_{d\in\{h,v\}} (\mathbf{D}_\mathbf{d}^\mathbf{T} Diag(\mathbf{m}_d \oslash (|\nabla_d\mathbf{l}| + \epsilon))\mathbf{D}_\mathbf{d}\right)\mathbf{t} = \mathbf{l}. \tag{6.16}$$

$\mathbf{I}$ is the unit matrix and $\mathbf{D_d}$ is the Toeplitz matrice.

By solving the optimization problem, the scene illumination $\mathbf{T}$ which is necessary to get the weight matrix $\hat{\mathbf{W}}$ is obtained.

## 6.3/ THE PROPOSED APPROACH

Our approach aims to tackle the day-night place recognition based on low-quality nighttime images and divides it into two parts: night vs. day place recognition and day vs. night place recognition. The first problem deals with selecting the image from the daytime reference images set that is the most similar as the nighttime query image. The second problem is the reverse: the query image is under daytime condition and set of reference images is under nighttime conditions.

As it is shown in Fig. 6.1, different feature selection processes are applied to solve the two kinds of day and night place recognition problem, and then the same place recognition framework (as shown in Fig. 3.1) is applied to retrieve the appropriate matching images. In detail, the feature similarity achievement approach consists in: 1) blurring the daytime images, 2) enhancing the nighttime images. These two stages are used to process both the day and nighttime raw images for the night vs. day place recognition. And for the day vs. night place recognition, only nighttime raw images are processed by the image enhancement and image inpainting methods. After applying the proposed methods to deal with the day and night images, the appropriately generated images are putted into CNN model to get their deep features. Through the deep features, the Euclidean distance are computed and nearest neighbor matching is used to achieve the place recognition purpose (general basic framework presented in section 3.1).

### 6.3.1/ THE FEATURE SELECTION PROCESS FOR NIGHT VS. DAY PLACE RECOGNITION

In this section, the feature similarity achievement approach to tackle night vs. day place recognition is described and illustrated in Fig. 6.3. Then, the dynamic objects removing system as proposed in Chapter 4 is applied in this approach to evaluate the impact of dynamic objects on the night and day place recognition, as shown in Fig. 6.4.

#### 6.3.1.1/ THE FEATURE SIMILARITY ACHIEVEMENT FOR NIGHT VS. DAY PLACE RECOGNITION APPROACH

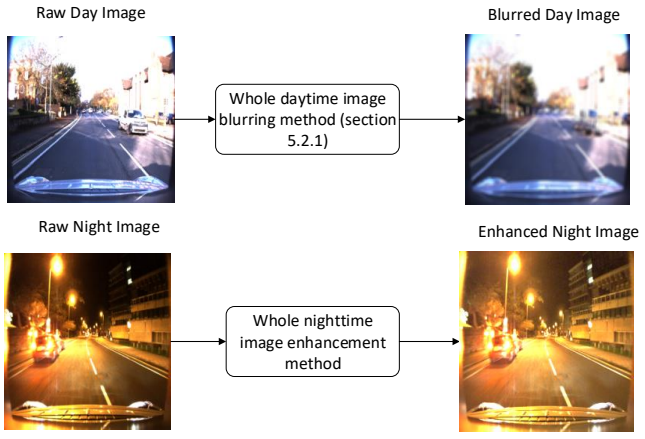



Figure 6.3: The process of feature similarity achievement approach for night vs. day place recognition problem.

Lots of researches [Anoosheh et al., 2019] [Porav et al., 2018] [Torii et al., 2015] proved that place recognition can have better performance if making the query and reference images having a similar appearance. Inspired by these approaches and by the observation that day images have better quality of features than night images, this chapter proposes a feature similarity achievement approach to make the features of the day and night images as similar as possible by: 1) decreasing the quality of the daytime images using image blurring method, 2) increasing the quality of the nighttime images using night image enhancement method.

Gaussian blur is used to blur the day images to make the features similar as captured in the night conditions, the effect can be seen in the blurred day image in Fig. 6.3.

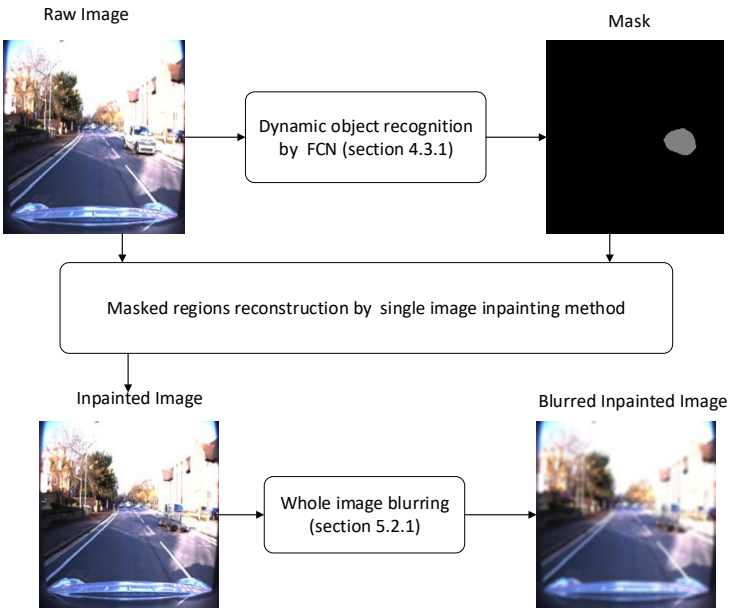
Images captured in non-uniform illuminated environments such as low-light and nighttime scenes may suffer from information loss in under-exposed regions, making the generated features lacking of distinguishability for place recognition. To increase the robustness of image features, it is important to reveal the information of under-exposed regions and to keep the information on well-exposed regions. The low-light image enhancement method described in section 6.2 is then used to enhance the image.

#### 6.3.1.2/ APPLYING THE OBJECTS REMOVING METHOD

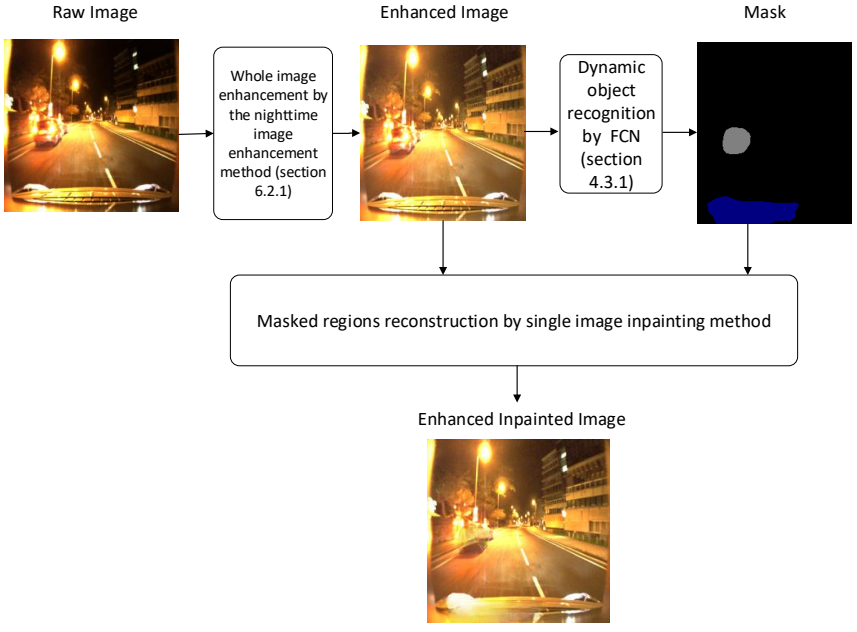
To recognize the dynamic objects in the day and night images as accurate as possible, the dynamic objects removing method should be applied in the images which have better quality. Therefore, for the daytime images, the dynamic objects removing is used firstly to generate the inpainted images in which the dynamic objects have been removed, then the image blurring method is used in the inpainted images to generate the inpainted blurred images, as shown in Fig 6.4 (a). For the nighttime images, the image enhancement method is used firstly to improve the quality of the images. Then, the dynamic objects are removed from the enhanced images by the FCN and the EdgeConnect methods and the inpainted enhanced images are obtained, as shown in Fig 6.4 (b).

As shown in Chapter 4, removing the dynamic objects in daytime images can improve the place recognition performance, the dynamic objects removing process is used in this chapter to see the influence of dynamic objects in day and night place recognition.

We recall that the dynamic objects removing method combines two steps: 1) recognizing the dynamic objects and generating the masks by semantic segmentation method, 2) removing the dynamic objects, and repairing the background information by image inpainting method based on the mask images. In this chapter, the single image inpainting method -EdgeConnect [Nazeri et al., 2019]- is chosen to repair the background information of the image based on the previous mask image generated by FCN [Long et al., 2015], which is the same as in Chapter 4.




(a) The dynamic objects removing and the image blurring process for day image.




(b) The dynamic objects removing and the image enhancement process for night image.

Figure 6.4: The proposed feature similarity achievement approach combined with dynamic objects removing approach for night vs. day place recognition problem. The top figure is the process for the day image, the bottom figure is the process for the night image.

### 6.3.2/ THE FEATURE SELECTION PROCESS FOR DAY VS. NIGHT PLACE RECOGNITION

The day vs. night place recognition problem selects the daytime image as query image and the nighttime images as the reference images. As the query image has good feature quality, the feature selection process focuses on improving the quality of the nighttime images, as shown in Fig. 6.5. Three kinds of enhanced images are generated by:

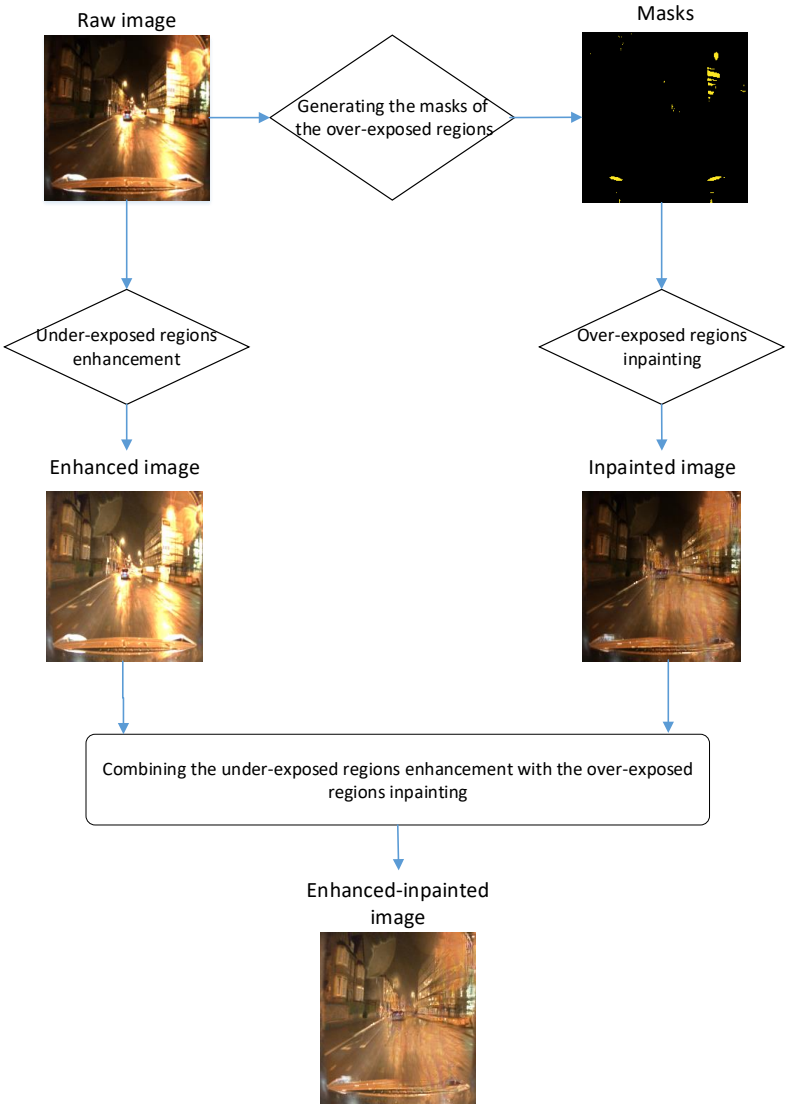



Figure 6.5: The feature selection process for day vs. night place recognition. The nighttime image enhancement and image inpainting methods are used and three kinds of enhanced images are generated.

- 1) repairing the hard light regions by image inpainting.

The hard light regions in the image may lose details because of the over-exposed problem. Finding a way to restore the details can contribute a lot to the building of robust feature representation. In this section, a simple method is proposed to recognize the over-exposed regions and generate the mask of each image. Having the hard over-exposed regions mask, the image inpainting approach can be used to restore the background information of the over-exposed regions in the mask.

Generating the mask $M$ basing on the illumination distribution of an image is a contribution of this chapter. The scene lightness map **L** is generated by the V channel of the HSV color representation model. The mask $M$ is a single channel image, and its values vary from 0 to 1. For the highly illuminated pixels (set of the pixels whose value is over 200), the value in the mask $M$ tends to be 1. For other pixels, the value in the mask $M$ is turned to be 0.

$$\mathbf{M}(x) = \begin{cases} 1, & \text{if } \mathbf{L}(x) \geq 200, \\ 0, & \text{if } \mathbf{L}(x) < 200. \end{cases} \tag{6.17}$$

- 2) enhancing the information of under-exposed regions of the night-time images, which is the same as the method (detailed in section 6.2.1) in night vs. day place recognition.

- 3) combining the image inpainting method to repair the hard light regions with image enhancement method to enhance the information of under-exposed regions of the night-time images.

### 6.3.3/ THE FRAMEWORK OF THE PACE RECOGNITION METHOD

Thorough the two feature selection processes, different kinds of daytime and nighttime images are generated. Then, the proposed images will be applied into the place recognition system, (as seen in Fig. 3.1). The features of both night-time images and the reference daytime images are generated by a featurization process. And then by comparing the features between one query image and each of the reference images using the metric of Euclidean distance, the closest matching reference image is seen as the matching result of the query image.

The featurization process in this paper is to get the deep features of the input images by a state-of-art ConvNet. To prove the robustness of our proposed methods, NIN_ImageNet [1] is used to extract deep features of input images. It is a 4 layer NIN (Network in Network) [Lin et al., 2013] model trained on Imagenet dataset.

The performance of the proposed approaches for place recognition will be reported, compared and analyzed in next section.

## 6.4/ RESULTS

This section presents the performance of the two proposed approaches (for night vs. day and day vs. night) and illustrates them with some image examples.

For the night vs. day place recognition, the performance of the feature similarity achievement approach can be seen in Table 6.1 and detailed in section 6.4.1.1. The best result of the proposed approach is compared to the state-of-art place recognition approaches. Based on the images generated by the feature similarity approach, the dynamic objects are removed to evaluate if further improvement in place recognition can be achieved. The results based on the dynamic objects removing method are shown in Table 6.2 and detailed in section 6.4.1.2.

For the day vs. night place recognition, Table 6.3 shows the performance using the proposed enhanced and inpainted images (see section 6.4.2).

### 6.4.1/ THE RESULTS OF NIGHT VS. DAY PLACE RECOGNITION

Fig. 6.6 shows examples of images generated by the proposed night vs. day place recognition approach. The feature similarity achievement method aims to achieve the balance of the quality of the nighttime and daytime images. Therefore, the daytime images are blurred by the image blurring method (as shown in the first line of Fig. 6.6 (b)), and the nighttime images are enhanced by the image enhancement method (as shown in the second, third and fourth lines of Fig. 6.6 (b)). The results in terms of place recognition using these images will be detailed in the first part of this section (6.4.1.1).

[1] https://github.com/dsys/nin-imagenet/tree/master/model

Applying the dynamic objects removal method after the feature similarity achievement method, new day and nighttime images are generated as illustrated in Fig. 6.6(c). The impact of these images on the night vs. day place recognition will be analyzed in the second part of this section (section 6.4.1.2).

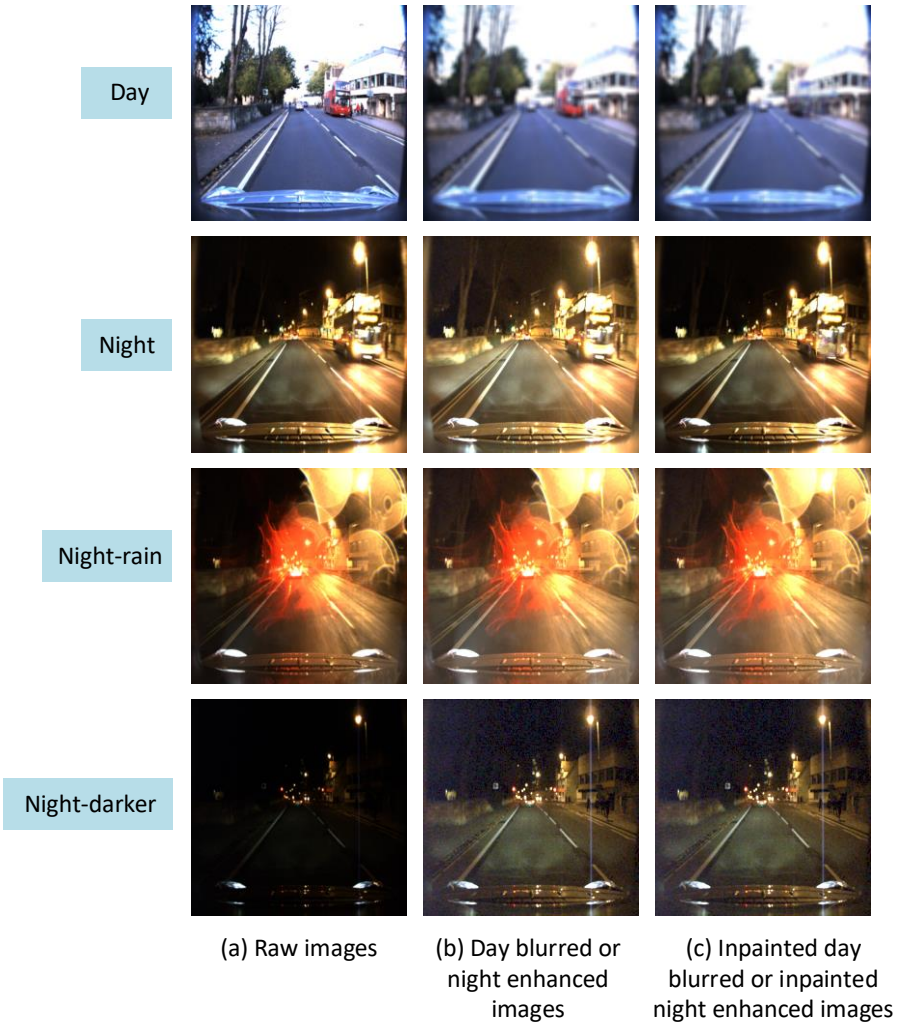



Figure 6.6: Raw images of day, night, night-rain and night-darker (first column). Blurred day image by image blurring method and enhanced night, night-rain and night-darker images by image enhancement method (second column). Inpainted day blurred image and inpainted enhanced night, night-rain and night-darker images generated by removing the dynamic objects in day blurred and three kinds of enhanced night images (third column).

Table 6.1: Night vs. day place recognition precision results using the proposed feature similarity achievement approach and comparison with two advanced methods of the literature (DenseVLAD and ToDayGAN).

| Precision(%) | Ours | | | | densevlad | ToDayGAN |
|---|---|---|---|---|---|---|
| | raw image | night-enhanced | day-blurred | night-enhanced day-blurred | | |
| night vs. day | 51.69 | 53.39 | 59.75 | **61.44** | 19.92 | 13.14 |
| night-rain vs. day | 22.88 | **25.85** | 23.73 | 25.42 | 10.88 | 11.72 |
| night-darker vs. day | 5.51 | 21.61 | 10.59 | **27.97** | 16.32 | 9.62 |

#### 6.4.1.1/ RESULTS OBTAINED WITH THE PROPOSED FEATURE SIMILARITY ACHIEVEMENT APPROACH

As shown in the second row of Fig. 6.6, the quality of the night images is normal in the night-time conditions. In the third row, night-rain images have lots of blurry regions due to the impact of the raindrops on the camera. The night-darker images, in the fourth row, have very bad quality and little information because of lacking of illumination. And compared to the daytime image, they all have worse quality. In this way, to balance the features of the day and night-time images and achieve the feature similarity, image blurring method are used in the daytime images and the damaging effect can be seen in the blurred day image in Fig. 6.6, at the same time, night-time images are enhanced and their information is increased.

As seen in the middle column of Fig. 6.6, the quality of all the three kinds of night-time images have been improved by image enhancement method. For the night images, the profile of the trees on the left side of the road can be seen more clearly in the enhanced night images than in the raw night images. Although the effect of enhancing is not very significant in the night-rain image, the top left part of the enhanced image has a weak improvement of its illumination. However, the enhancing effect in the night-darker image is very strong and the buildings appear on the right side of the enhanced image while they were unsuspected in the raw image.

As the image enhancement method improves the quality of the night-time images, the place recognition performance has been improved when the enhanced night-time images are used as the query images, as shown in the third column in Table 6.1. Thanks to the huge improvement of the quality of the night-darker images, the place recognition results

by applying enhanced night-darker images have better improvement than the results by applying the other two kinds of images.

In the fourth column of Table 6.1, results show that blurring the reference images can also improve the place recognition performance. This is an inspiring result because it indicates a new direction of place recognition, that damaging the quality of images can take better performance in return. This is perhaps because the query night image is lacking details, and the clear details of daytime images may be a burden and mislead the feature matching process of the query and reference images.

The results in the fifth column of Table 6.1 are the place recognition results obtained by using both the enhanced night query image and the day blurred reference image, which realizes the proposed feature similarity achievement approach and gets the best place recognition performance in the night-day and night-darker-day sequences compared to the other methods. The result in the night-rain-day sequence is a little worse than the result obtained using the enhanced night images only, but better than using the raw images or the blurred images only. This is perhaps due to the influence of the raindrop regions but it needs more research to make it clear.

The results of our methods are better than densevlad and ToDayGAN. For the night-darker sequence, ToDayGAN and densevlad have better performance than the results by directly putting the raw images into the ConvNet. It illustrates the robustness of densevlad and ToDayGAN facing very little information query image, but our proposed methods have better performance.

In fact, ToDayGAN translates the nighttime images into daytime ones firstly. Then it uses the method of densevlad to match real daytime reference images with the synthesized daytime query image. In contrary, densevlad uses the raw night and day images directly in our experiments. The worse results of ToDayGAN in night-day and night-darker-day sequences show that the image-to-image translation network in ToDayGAN performs not very well for the low-quality nighttime images.

This is the drawback of the recent researches which apply image-to-image translation network in day and night place recognition. The networks are well-trained for certain datasets, but when changing the dataset, it may have worse performance. But our proposed method is very simple to use and adjust for day and night place recognition as the

quality of the nighttime images is normally worse than daytime images.

#### 6.4.1.2/ RESULTS BY THE PROPOSED FEATURE SIMILARITY ACHIEVEMENT APPROACH COMBINED WITH DYNAMIC OBJECTS REMOVING APPROACH

Table 6.2: Place recognition precision results using the proposed feature similarity achievement approach combined with dynamic objects removing approach (the precision obtained without the dynamic obstacles removing step is as the baseline in the second column).

| Precision(%) | night-enhanced/ day-blurred **(baseline)** | night-enhanced/ day-blurred-inpainted | night-enhanced-inpainted/ day-blurred | night-enhanced-inpainted/ day-blurred-inpainted |
|---|---|---|---|---|
| night vs. day | **61.44** | **64.41** | **61.86** | **63.98** |
| night-rain vs. day | **25.42** | 22.46 | 25.42 | 22.46 |
| night-darker vs. day | **27.97** | 27.97 | 27.54 | 27.97 |

Since blurring the daytime image and enhancing the night-time image can improve the day and night place recognition, more experiments are done to evaluate if removing the dynamic objects in the blurred day images and enhanced nighttime images can make further improvement in the place recognition performance. Selecting the best result of the feature similarity achievement approach as a baseline (result by using both the night-enhanced and day-blurred images), the experiments are set in three configurations: adding dynamic objects removal step in 1) both the night-enhanced and day-blurred images, 2) only the night-enhanced images and 3) only the day-blurred images.

From the results of the night vs. day sequence in the second row in Table 6.2, it can be seen that using the three configurations can all improve the place recognition accuracy (2.54% increase by configuration 1, 0.42% increase by configuration 2 and 2.97% increase by configuration 3 compared to precision 61.44% as baseline). But we can see that the best result is obtained by only removing the dynamic objects from the blurred daytime images (precision 64.41%), and adding the dynamic objects removal step in the enhanced night images then decreases the result (precision 63.98%). As illustrated in the images shown in Fig. 6.6, the bus in the day image is well removed and the back-

ground is inpainted in the inpainted blurred image. Even though there is an artifact in the inpainted blurred image, it does not matter because that the whole image is blurry. Unlike the daytime image, the bus in the night image is not well recognized perhaps because its profile is not very clear in the night conditions, thus resulting in the terrible removing performance in the inpainted enhanced night image. That is why removing the dynamic objects from the blurred daytime images enhances the place recognition performance while removing them from the enhanced nighttime images decreases the performance.

For the night-rain vs. day sequence and night-darker vs. day sequence, the results are dissatisfied. For the night-rain-day sequence, removing the dynamic objects from only the enhanced night images do not have any influence on the place recognition performance (precision 25.42%, the same as the baseline precision), removing them from only the blurred day images is worse more - damaging the place recognition performance by 0.96% decrease of accuracy compared to the baseline, and removing them from both the blurred day images and enhanced night images also has a 0.96% precision decrease. For the night-darker-day sequence, inpainting both day and night images seems to have nearly no impact on the place recognition performance.

It is not hard to explain why both the inpainted night-rain and inpainted night-darker images do not have any effect. The night-rain images are too blurry and the night-darker images have little information, the FCN for the dynamic objects removing method cannot recognize any dynamic objects in the images, and therefore their dynamic objects are not removed at all. It can be concluded that for the nighttime images which have little information, it is not necessary to use the proposed dynamic objects removing method.

It is very confusing that taking the same day images as reference images, when the nighttime query images are changing, the effect of removing the objects in daytime images on place recognition can be different. For the night vs. day sequence, it improves the place recognition performance; for the night-rain vs. day sequence, it damages the place recognition performance; for the night-darker vs. day sequence, it nearly has no impact on place recognition. Facing this result, we can just suppose that the quality of the query image plays an important role. If the query images have normal nighttime quality, which means they are obtained in enough nighttime illumination and does not obtained in severe weather conditions, the presence of cars, buses, pedestrians may damage the place recognition performance, and applying the proposed method to remove them can improve

the performance.

### 6.4.2/ THE RESULTS OF THE DAY VS. NIGHT PLACE RECOGNITION

In this section, examples of raw-query sequences and the three kinds of enhanced images are given to illustrate the visual effects of the proposed methods. The results of place recognition based on those four kinds of images are given to show the advanced performance of the proposed methods compared to the two state-of-art place recognition methods.

Table 6.3: Place recognition precision results using the proposed approaches and comparison with two advanced methods of the literature (densevlad and ToDayGAN).

| Precision(%) | Ours | | | | densevlad | TodayGAN |
|---|---|---|---|---|---|---|
| | raw | enhanced | inpainted | enhance_inpainted | | |
| day V.S night | 51.69 | 58.90 | 61.02 | 64.83 | 21.61 | 10.17 |
| day V.S night-rain | 24.15 | 28.39 | 24.58 | 26.69 | 13.98 | 11.86 |
| day V.S night-darker | 21.61 | 44.92 | 26.27 | 53.81 | 17.57 | 9.32 |

#### 6.4.2.1/ THE PLACE RECOGNITION RESULTS OF THE PROPOSED METHODS BASED ON THE DAY AND NIGHT IMAGES

(a) Raw (b) Enhanced (c) Inpainted (d) Enhance-inpainted

Figure 6.7: Examples of night query images. The under-exposed region (top left of a) is well enhanced and showed in b and d. The over-exposed regions (middle and top right parts of a) are inpainted and showed in c and d.

The night images are blurry but with normal night-time brightness, as shown in Fig. 6.7. There are lots of over-exposed regions in the image in this light condition. The top-left region of the raw night image (Fig. 6.7 (a)) is under-exposed, and is well enhanced as

one can see in the enhanced image (Fig. 6.7 (b)) . And the top-right region of the raw night image is over-exposed, the image of Fig. 6.7 (c) shows that the over-exposed region is well recognized and restored (Fig. 6.7 (c)). But the restoring performance is not perfect and brings some noisy information in this region because the image inpainting method is insufficient when facing large portion missing problem. In contrast, image enhancement performs perfect results in enhancing the brightness of the whole image, as shown in Fig. 6.7 (d).

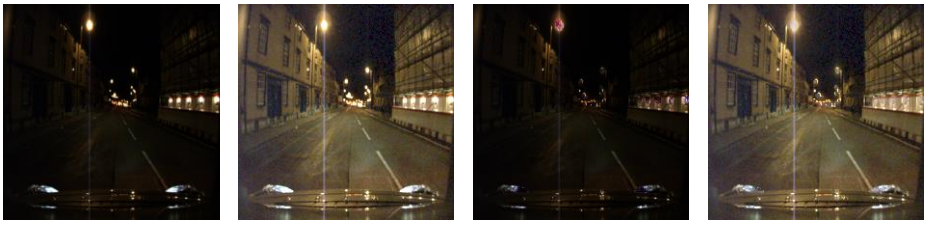
Therefore, enhancing the lightness of under-exposed regions (precision 72.88%) and recognizing and restoring the over-exposed regions (precision 67.80%) can both improve the day and night place recognition, but the result of combining them (precision 72.03%) performs worse than using only image enhancement, as can be seen in the first row of Table 6.3.

As ToDayGAN can not generate high-quality daytime images based on this night sequence, it damages the place recognition performance and therefore performs worse than densevlad. Although densevlad performs well, the proposed method performs better thanks to the increasing of feature information.

#### 6.4.2.2/ THE PLACE RECOGNITION RESULTS OF THE PROPOSED METHODS BASED ON THE DAY AND NIGHT-RAIN IMAGES


Figure 6.8: Examples of night-rain query images. The under-exposed region (left part of a) is well enhanced and shown in b and d. The over-exposed region (right part of a) is inpainted and shown in c and d.

Images captured at night during rain have very bad quality because of the impact of the raindrops on the camera, as can be seen in the raw image of Fig. 6.8. The over-exposed regions are highly increased because of the reflection of light during rain. So restoring



(a) Raw (b) Enhanced (c) Inpainted (d) Enhance-inpainted

Figure 6.9: Examples of night-darker query images. The under-exposed region (whole of a) is well enhanced and shown in b and d.

those over-exposed regions increases the key information of images ((c) and (d) in Fig. 6.8) and plays an important role in improving the place recognition performance. At the same time, enhancing the lightness of under-exposed regions ((b) in Fig. 6.8) can always work in day and night place recognition. Therefore, both enhancing the images and inpainting the hard light regions improve the place recognition performance but combining these two methods improves more the results, as seen in the second row of Table 6.3.

Compared to our proposed methods, it is difficult for the night-to-day translation model to translate the nighttime images because of lacking some key information of the image. Therefore, it is not strange that ToDayGAN generates bad quality daytime images and has bad place recognition performance. As for densevlad, Table 6.3 shows that it can not perform well under such huge appearance variations between day and night-rain images.

#### 6.4.2.3/ THE PLACE RECOGNITION RESULTS OF THE PROPOSED METHODS BASED ON THE DAY AND NIGHT-DARKER (VERY LOW ILLUMINATION) IMAGES

The image enhancement method has the most advantage when facing the images with little illumination. So it can be seen from Fig. 6.9 that the enhanced image reveals much more information than the raw-image. In this sequence, the raw images have little luminance and each of the over-exposed regions represents only a small portion of the whole image. The image inpainting method is very good at recovering small missing regions. So using the EdgeConnect to restore the background information of the over-exposed regions can increase the image features without bringing more noise.

That is the reason why inpainting the over-exposed regions, and combining it with enhancing the under-exposed regions together can both improve the place recognition per-

formance. But we note that enhancing the brightness of the under-exposed regions is more important in this somber night condition. These statements are confirmed by the results shown in the third row of Table 6.3.

Table 6.3 shows also that ToDayGAN and densevlad methods have a worse performance again because of lacking image information.

## 6.5/ CONCLUSION AND FUTURE WORKS

In this chapter we divided the day and night place recognition problem into two parts and proposes two different methods to tackle them. From the evaluation results, it can be concluded that the feature similarity approach - blurring the daytime images and enhancing the night-time images - is very effective for night vs. day place recognition. For the day vs. night place recognition, applying the image inpainting method and image enhancement to improve the robustness of nighttime image features is enough to get good results. Facing the low-quality of night-time images, the proposed methods may be a better choice than using the night-to-day image translation network. Even though, applying the night-to-day image translation network in day and night place recognition is till a promising research direction. In our future work, we plan to train a model that can translate the nighttime image into day no matter what kind of nighttime dataset is.

Furthermore, the dynamic objects removing method proposed in Chapter 4 may further improve the place recognition performance based on using images dealt with our feature similarity approach. But this may affect only in using the night images that have a normal quality as the query image. Therefore, our future work is to find more appropriate nighttime dataset to make deep analysis and conclusions.

# III

# Conclusions and Future Works

# 7

# CONCLUSIONS AND FUTURE WORKS

## 7.1/ CONCLUSIONS

As described in Chapter 1, place recognition is an important step for the image-based localization tasks. This thesis aimed to improve the performance of vehicle localization based on scene perception and focused on the place recognition part. As feature representation and matching determines the success for recognizing a place, the methods to generate robust image features were proposed in this thesis.

Firstly, removing dynamic object for place recognition is proposed in Chapter 4. If the image of a place is captured in busy traffic conditions, there will be many vehicles and pedestrians. The severe changing in presence of pedestrians or vehicles can damage the robustness of features. We therefore proposed to remove them in our method to generate robust features. Semantic segmentation is used to detect and mask the dynamic objects, and the masked dynamic objects can be seen as missing regions that have been reconstructed according to the background regions using image inpainting. FCN, one of the simple semantic segmentation methods is chosen in this method and two image inpainting methods -single image inpainting and sequence image inpainting- are used to generate two kinds of inpainted images. In the place recognition system, CNN model is chosen to generate the image features and the nearest neighbour searching is used to find the best matching place.

Different experiments were designed to analyze the influence of the inpainted images. First of all, the place recognition performance is compared based on the three different input images: raw image, single-inpainted image and seq-inpainted image, which proved the effectiveness of the proposed inpainted images. Then four kinds of CNN models

were used to generate different kinds of deep features of these input images. The results showed that the choice of CNN model has no impact on the conclude: the place recognition performance of using inpainted images is always better than using raw images. Experiments were also done by reversing the query and reference images of the same dataset, for example, the summer vs. winter place recognition is reversed as winter vs. summer place recognition based on the same dataset. It proved that the quality of the query image determines the place recognition performance. Finally, the results of place recognition based on both inpainted query and reference sequences and either the inpainted query or inpainted reference sequences were compared, which concluded that using only the inpainted reference image can guarantee the good place recognition performance. The proposed method has proved its effectiveness when compared to two advanced place recognition approaches: DenseVLAD and OLO.

Then, reducing image noise based place recognition is proposed in Chapter 5. This is inspired by the work of the Chapter 4. We proposed to reduce the noisy information (dynamic objects) instead of removing them. For that, image blurring method is used in this chapter. We considered both blurring the dynamic objects recognized by FCN in Chapter 4, and the whole image.

The experiments were designed based on the quality of the images. Firstly, for the high quality daytime images, the place recognition performance based on both the whole blurred images and the blurred-fcn images was tested. The results showed that for the whole blurred images, the place recognition based on only the reference blurred images can have better performance under the condition that the reference image have better quality than the query image. As for the blurred-fcn images, the results were not well. It seems that using both the blurred query and reference images or neither of them have no impact on place recognition and can not always improve the performance. The dynamic objects blurring method may improve the place recognition performance only in busy traffic conditions.

For nighttime place recognition, only the whole blurred images were used for experiments because cars cannot be obviously seen in this condition. Then, the conclusion is that blurring the whole image can improve the place recognition performance under low quality query vs. good quality reference images.

As the method to measure the image quality has not proposed in this thesis yet, the night

vs. day place recognition based on the whole blurred images was designed in this chapter. As nighttime images have a bad quality and daytime images have a good quality, a manual designed low quality query vs. good quality reference images conditions can be used to further prove the above conclusion. For the daytime condition, the proposed methods were compared with DenseVLAD and OLO methods. And for the nighttime or day vs. night place recognition, the proposed methods were compared with only DenseVLAD. The results showed that the proposed methods have better place recognition performance in most conditions.

Thirdly, the approach of day and night place recognition is proposed in Chapter6. The day and night place recognition can be divided into night vs. day and day vs. night place recognition because the quality of the query image influences the place recognition performance, as proved in the first two contributions. For the night vs. day place recognition, two methods were proposed. The first one is the feature similarity achievement, consisting of enhancing the nighttime images and blurring the daytime images. Then, we combined this method with the dynamic objects removing methods proposed in Chapter 4. When evaluating the two methods on the three dataset, the results proved that the feature similarity achievement method is better than the DenseVLAD and ToDayGAN methods. As for combing it with the dynamic objects removing method, it can improve the place recognition performance in the normal night conditions as cars in this condition are much more obvious and may damage the place recognition performance. But for the rainy night condition or very dark night condition, this method seems have no impact on the place recognition performance.

For the day vs. night place recognition, as the daytime query images are in good quality condition, it is better to keep them unchanged. Methods were proposed to improve the quality of nighttime image by enhancing its under-exposed regions and inpainting the over-exposed regions according to the illumination. When compared with DenseVLAD and ToDayGAN methods, this method also showed its effectiveness in the day vs. night place recognition.

## 7.2/ FUTURE WORKS

Although the proposed approaches have shown good performances in place recognition, they could be developed and improved in the future works.

**1)** Image inpainting methods were used in Chapter 4. Even there exist advanced image inpainting methods, artifacts and some noisy information are more or less brought in the inpainting results. If improving the image inpainting performance and removing perfectly the dynamic objects, the place recognition should be improved further. To achieve this, new image inpainting method could be proposed in the future by either designing a new image inpainting framework or training the existing image inpainting network on the specific datasets for the place recognition task.

**2)** In chapter 5, the quality of an image is an important factor to be considered, but there exist no metric that can be used to evaluate the image quality and it can be only estimated visually. Finding a measure to evaluate precisely the image quality will be a continuous focus point in our future work.

**3)** More methods to reduce the noisy information of an image need to be explored. The most commonly image blurring method is Gaussian blurring, which is used in this thesis. There are many other image blurring methods, such as median filter or bilateral filter, that can be used. Beside image blurring methods, methods, as long as they are able to reduce the image noise, can be considered to develop the work of Chapter 5.

**4)** As for the day and night place recognition problem, although the approaches proposed in Chapter 6 have better performance than the night to day image translation methods under low-quality image conditions, the night to day translation networks are still a promising direction for place recognition. The limit of these networks is that they can only have good performance in specific datasets which are similar to the training datasets. As the training time is very consuming, it is not possible to train the network for all possible conditions. Therefore, the transportability of the network is an important point to be developed. Then, the night to day translation based on low-quality images is also difficult for the network as the features of the low-quality images are too limited to make the network well learned. If these difficul-

ties can be tackled, the place recognition performance based on night to day image translation network can be improved drastically.

**5)** Combining all the contributions in this thesis as an end-to-end network could be a worthy work in the future. Given an image as the input, it can be automatically chosen to enter the image blurring, image enhancing or image inpainting process, and then the pre-processed image is inputted into a CNN model to get the deep features. This end-to-end network could be suitable for place recognition under many conditions.

# BIBLIOGRAPHY


**[Abolfazli Esfahani et al., 2019]** Abolfazli Esfahani, M., Wu, K., Yuan, S., et Wang, H. (2019). **DeepDSAIR: Deep 6-dof camera relocalization using deblurred semantic-aware image representation for large-scale outdoor environments**. *Image and Vision Computing*, 89:120 – 130.

**[Agrawal et al., 2008]** Agrawal, M., Konolige, K., et Blas, M. R. (2008). **Censure: Center surround extremas for realtime feature detection and matching**. In *Computer Vision – ECCV*, pages 102–115, Berlin, Heidelberg. Springer Berlin Heidelberg.

**[Alahi et al., 2012]** Alahi, A., Ortiz, R., et Vandergheynst, P. (2012). **Freak: Fast retina keypoint**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–517.

**[Amadeo et al., 2016]** Amadeo, M., Campolo, C., et Molinaro, A. (2016). **Information-centric networking for connected vehicles: a survey and future perspectives**. *IEEE Communications Magazine*, 54(2):98–104.

**[Anati et al., 2009]** Anati, R., et Daniilidis, K. (2009). **Constructing topological maps using markov random fields and loop-closure detection**. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

**[Anoosheh et al., 2018]** Anoosheh, A., Agustsson, E., Timofte, R., et Van Gool, L. (2018). **Combogan: Unrestrained scalability for image domain translation**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

**[Anoosheh et al., 2019]** Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., et Gool, L. V. (2019). **Night-to-day image translation for retrieval-based localization**. In *International Conference on Robotics and Automation (ICRA)*, pages 5958–5964.

**[Arandjelovic et al., 2016]** Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., et Sivic, J. (2016). **Netvlad: Cnn architecture for weakly supervised place recognition**. In

*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[Arandjelović et al., 2012]** Arandjelović, R., et Zisserman, A. (2012). **Three things everyone should know to improve object retrieval**. In *IEEE Conference on Computer Vision and Pattern Recognition*.

**[Arroyo et al., 2014a]** Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., Yebes, J. J., et Bronte, S. (2014a). **Fast and effective visual place recognition using binary codes and disparity information**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3089–3094.

**[Arroyo et al., 2014b]** Arroyo, R., Alcantarilla, P. F., Bergasa, L. M., Yebes, J. J., et Gámez, S. (2014b). **Bidirectional loop closure detection on panoramas for visual navigation**. In *IEEE Intelligent Vehicles Symposium Proceedings*, pages 1378–1383.

**[Arth et al., 2015]** Arth, C., Pirchheim, C., Ventura, J., Schmalstieg, D., et Lepetit, V. (2015). **Instant outdoor localization and slam initialization from 2.5d maps**. *IEEE Transactions on Visualization  Computer Graphics*, 21(11):1309–1318.

**[Ascani et al., 2008]** Ascani, A., Frontoni, E., Mancini, A., et Zingaretti, P. (2008). **Feature group matching for appearance-based localization**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3933–3938.

**[Badino et al., 2011]** Badino, H., Huber, D., et Kanade, T. (2011). **Visual topometric localization**. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 794–799.

**[Badino et al., 2012]** Badino, H., Huber, D., et Kanade, T. (2012). **Real-time topometric localization**. In *IEEE International Conference on Robotics and Automation*, pages 1635–1642.

**[Badrinarayanan et al., 2017]** Badrinarayanan, V., Kendall, A., et Cipolla, R. (2017). **Segnet: A deep convolutional encoder-decoder architecture for image segmentation**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495.

**[BAI et al., 2018]** BAI, D., WANG, C., ZHANG, B., YI, X., et YANG, X. (2018). **Cnn feature boosted seqslam for real-time loop closure detection**. *Chinese Journal of Electronics*, 27:488–499(11).

**[Bai et al., 2018]** Bai, D., Wang, C., Zhang, B., Yi, X., et Yang, X. (2018). **Sequence searching with CNN features for robust and fast visual place recognition**. *Computers  Graphics*, 70:270 – 280.

**[Baluja, 1996]** Baluja, S. (1996). **Evolution of an artificial neural network based autonomous land vehicle controller**. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 26(3):450–463.

**[Bay et al., 2006a]** Bay, H., Tuytelaars, T., et Van Gool, L. (2006a). **Surf: Speeded up robust features**. In *Computer Vision – ECCV*, pages 404–417. Springer Berlin Heidelberg.

**[Bay et al., 2006b]** Bay, H., Tuytelaars, T., et Van Gool, L. (2006b). **SURF: Speeded Up Robust Features**. In *European Conference on Computer Vision (ECCV)*, pages 404–417, Berlin, Heidelberg.

**[Bertalmio et al., 2000]** Bertalmio, M., Sapiro, G., Caselles, V., et Ballester, C. (2000). **Image inpainting**. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, page 417–424, USA.

**[Booij et al., 2009]** Booij, O., Zivkovic, Z., et Kröse, B. (2009). **Efficient data association for view based slam using connected dominating sets**. *Robotics and Autonomous Systems*, 57(12):1225–1234.

**[Bradley et al., 2005]** Bradley, D. M., Patel, R., Vandapel, N., et Thayer, S. M. (2005). **Real-time image-based topological localization in large outdoor environments**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3670–3677.

**[Bresson et al., 2017]** Bresson, G., Alsayed, Z., Yu, L., et Glaser, S. (2017). **Simultaneous localization and mapping: A survey of current trends in autonomous driving**. *IEEE Transactions on Intelligent Vehicles*, 2(3):194–220.

**[Broggi et al., 2013]** Broggi, A., Buzzoni, M., Debattisti, S., Grisleri, P., Laghi, M. C., Medici, P., et Versari, P. (2013). **Extensive tests of autonomous driving technologies**. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1403–1415.

**[Calonder et al., 2010]** Calonder, M., Lepetit, V., Strecha, C., et Fua, P. (2010). **Brief: Binary robust independent elementary features**. In *Computer Vision – ECCV*, pages 778–792. Springer Berlin Heidelberg.

**[Chang et al., 2019]** Chang, Y.-L., Yu Liu, Z., et Hsu, W. (2019). **Vornet: Spatio-temporally consistent video inpainting for object removal**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

**[Chapoulie et al., 2013]** Chapoulie, A., Rives, P., et Filliat, D. (2013). **Appearance-based segmentation of indoors/outdoors sequences of spherical views**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1946–1951.

**[Chen et al., 2015a]** Chen, C., Seff, A., Kornhauser, A., et Xiao, J. (2015a). **Deepdriving: Learning affordance for direct perception in autonomous driving**. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

**[Chen et al., 2015b]** Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., et Yuille, A. L. (2015b). **Semantic image segmentation with deep convolutional nets and fully connected crfs**.

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

**[Chen et al., 2014]** Chen, Z., Lam, O., Jacobson, A., et Milford, M. (2014). **Convolutional neural network-based place recognition**. In Chen, C., editor, *Proceedings of the 16th Australasian Conference on Robotics and Automation*, pages 1–8. Australian Robotics and Automation Association Inc., Australia.

**[Chen et al., 2018]** Chen, Z., Liu, L., Sa, I., Ge, Z., et Chli, M. (2018). **Learning context flexible attention model for long-term visual place recognition**. *IEEE Robotics and Automation Letters*, 3(4):4015–4022.

**[Chen et al., 2015]** Chen, Z., Lowry, S., Jacobson, A., Ge, Z., et Milford, M. (2015). **Distance metric learning for feature-agnostic place recognition**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2556–2563.

**[Chen et al., 2017b]** Chen, Z., Maffra, F., Sa, I., et Chli, M. (2017b). **Only look once, mining distinctive landmarks from convnet for visual place recognition**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9–16.

**[Choi et al., 2018]** Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., et Choo, J. (2018). **Stargan: Unified generative adversarial networks for multi-domain image-to-image translation**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[Churchill et al., 2013]** Churchill, W., et Newman, P. (2013). **Experience-based navigation for long-term localisation**. *The International Journal of Robotics Research*, 32(14):1645–1661.

**[Colombo et al., 2012]** Colombo, A., et Del Vecchio, D. (2012). **Efficient algorithms for collision avoidance at intersections**. In *Proceedings of the 15th ACM International Conference on Hybrid Systems: Computation and Control*, HSCC '12, page 145–154.

**[Cummins et al., 2008]** Cummins, M., et Newman, P. (2008). **Fab-map: Probabilistic localization and mapping in the space of appearance**. *The International Journal of Robotics Research*, 27(6):647–665.

**[Ding et al., 2019]** Ding, D., Ram, S., et Rodríguez, J. J. (2019). **Image inpainting using nonlocal texture matching and nonlinear filtering**. *IEEE Transactions on Image Processing*, 28(4):1705–1719.

**[Doersch et al., 2015]** Doersch, C., Singh, S., Gupta, A., Sivic, J., et Efros, A. A. (2015). **What makes paris look like paris?** *Commun. ACM*, 58(12):103–110.

**[Dong et al., 2019]** Dong, J., Yin, R., Sun, X., Li, Q., Yang, Y., et Qin, X. (2019). **Inpainting of remote sensing sst images with deep convolutional generative adversarial network**. *IEEE Geoscience and Remote Sensing Letters*, 16(2):173–177.

**[ElHarrouss et al., 2019]** ElHarrouss, O., Almaadeed, N., Al-Máadeed, S., et Akbari, Y. (2019). **Image inpainting: A review**. *CoRR*, abs/1909.06399.

**[Eureka, ]** Eureka. **E! 45: Programme For a European Traffic System With Highest Efficiency and Unprecedented Safety.** Accessed: May 19, 2019. [online]. http://www.eurekanetwork.org/project/id/45, year = 1987.

**[Feng et al., 2016]** Feng, Y., Fan, L., et Wu, Y. (2016). **Fast localization in large-scale environments using supervised indexing of binary features**. *IEEE Transactions on Image Processing*, 25(1):343–358.

**[Ružić et al., 2015]** Ružić, T., et Pižurica, A. (2015). **Context-aware patch-based image inpainting using markov random field modeling**. *IEEE Transactions on Image Processing*, 24(1):444–456.

**[Fröhlich et al., 2013]** Fröhlich, B., Rodner, E., et Denzler, J. (2013). **Semantic segmentation with millions of features: Integrating multiple cues in a combined random forest approach**. In *Computer Vision – ACCV*, pages 218–231.

**[Garcia-Fidalgo et al., 2015]** Garcia-Fidalgo, E., et Ortiz, A. (2015). **Vision-based topological mapping and localization methods: A survey**. *Robotics and Autonomous Systems*, 64:1–20.

**[Gaspar et al., 2000]** Gaspar, J., Winters, N., et Santos-Victor, J. (2000). **Vision-based navigation and environmental representations with an omnidirectional camera**. *IEEE Transactions on Robotics and Automation*, 16(6):890–898.

**[Gerla et al., 2014]** Gerla, M., Lee, E., Pau, G., et Lee, U. (2014). **Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds**. In *IEEE World Forum on Internet of Things (WF-IoT)*, pages 241–246.

**[Glover et al., 2010]** Glover, A. J., Maddern, W. P., Milford, M. J., et Wyeth, G. F. (2010). **Fab-map + ratslam: Appearance-based slam for multiple times of day**. In *IEEE International Conference on Robotics and Automation*, pages 3507–3512.

**[Gu et al., 2018]** Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et Chen, T. (2018). **Recent advances in convolutional neural networks**. *Pattern Recognition*, 77:354–377.

**[Guerrero et al., 2018]** Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M., Dickie, D., Wardlaw, J., et Rueckert, D. (2018). **White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks**. *NeuroImage: Clinical*, 17:918 – 934.

**[Guo et al., 2018]** Guo, Q., Gao, S., Zhang, X., Yin, Y., et Zhang, C. (2018). **Patch-based image inpainting via two-stage low rank approximation**. *IEEE Transactions on Visualization and Computer Graphics*, 24(6):2023–2036.

**[Guo et al., 2017]** Guo, X., Li, Y., et Ling, H. (2017). **Lime: Low-light image enhancement via illumination map estimation**. *IEEE Transactions on Image Processing*, 26(2):982–993.

**[Hafner et al., 2013]** Hafner, M. R., Cunningham, D., Caminiti, L., et Del Vecchio, D. (2013). **Cooperative collision avoidance at intersections: Algorithms and experiments**. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1162–1175.

**[Han et al., 2019]** Han, X., Wu, Z., Huang, W., Scott, M. R., et Davis, L. S. (2019). **Compatible and diverse fashion image inpainting**.

**[Hausler et al., 2019]** Hausler, S., Jacobson, A., et Milford, M. (2019). **Filter early, match late: Improving network-based visual place recognition**. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

**[He et al., 2017]** He, K., Gkioxari, G., Dollar, P., et Girshick, R. (2017). **Mask r-cnn**. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

**[He et al., 2016]** He, K., Zhang, X., Ren, S., et Sun, J. (2016). **Deep residual learning for image recognition**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[He et al., 2006]** He, X., Zemel, R. S., et Mnih, V. (2006). **Topological map learning from outdoor image sequences**. *Journal of Field Robotics*, 23(11-12):1091–1104.

**[Hernan Badino et al., 2011]** Hernan Badino, D. H., et Kanade, T. (2011). **The CMU visual localization data set**. *http://3dvis.ri.cmu.edu/data-sets/localization*.

**[Hinton et al., 2012]** Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., et Salakhutdinov, R. R. (2012). **Improving neural networks by preventing co-adaptation of feature detectors**.

**[Hou et al., 2015]** Hou, Y., Zhang, H., et Zhou, S. (2015). **Convolutional neural network-based image representation for visual loop closure detection**. In *IEEE International Conference on Information and Automation*, pages 2238–2245.

**[Ignatov et al., 2018]** Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., et Van Gool, L. (2018). **Wespe: Weakly supervised photo enhancer for digital cameras**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

**[Isola et al., 2017]** Isola, P., Zhu, J.-Y., Zhou, T., et Efros, A. A. (2017). **Image-to-image translation with conditional adversarial networks**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[Jin et al., 2015]** Jin, K. H., et Ye, J. C. (2015). **Annihilating filter-based low-rank hankel matrix approach for image inpainting**. *IEEE Transactions on Image Processing*, 24(11):3498–3511.

**[Johnson et al., 2016a]** Johnson, J., Alahi, A., et Fei-Fei, L. (2016a). **Perceptual losses for real-time style transfer and super-resolution**. In *Computer Vision – ECCV*, pages 694–711, Cham.

**[Johnson et al., 2016b]** Johnson, J., Alahi, A., et Fei-Fei, L. (2016b). **Perceptual losses for real-time style transfer and super-resolution**. In Leibe, B., Matas, J., Sebe, N., et Welling, M., editors, *Computer Vision – ECCV*, pages 694–711. Springer International Publishing.

**[Junqiu Wang et al., 2006]** Junqiu Wang, Hongbin Zha, et Cipolla, R. (2006). **Efficient topological localization using orientation adjacency coherence histograms**. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 2, pages 271–274.

**[Kawai et al., 2016]** Kawai, N., Sato, T., et Yokoya, N. (2016). **Diminished reality based on image inpainting considering background geometry**. *IEEE Transactions on Visualization and Computer Graphics*, 22(3):1236–1247.

**[Kendall et al., 2015]** Kendall, A., Grimes, M., et Cipolla, R. (2015). **Posenet: A convolutional network for real-time 6-dof camera relocalization**. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

**[Kim et al., 2019]** Kim, D., Woo, S., Lee, J.-Y., et Kweon, I. S. (2019). **Deep video inpainting**. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[Kim et al., 2017]** Kim, H. J., Dunn, E., et Frahm, J. (2017). **Learned contextual feature reweighting for image geo-localization**. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3251–3260.

**[Konolige et al., 2010]** Konolige, K., Bowman, J., Chen, J., Mihelich, P., Calonder, M., Lepetit, V., et Fua, P. (2010). **View-based maps**. *The International Journal of Robotics Research*, 29(8):941–957.

**[Kosecka et al., 2003]** Kosecka, J., Liang Zhou, Barber, P., et Duric, Z. (2003). **Qualitative image based localization in indoors environments**. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–II.

**[Koutník et al., 2013]** Koutník, J., Cuccu, G., Schmidhuber, J., et Gomez, F. (2013). **Evolving large-scale neural networks for vision-based reinforcement learning**. In *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation*, GECCO '13, page 1061–1068.

**[Krizhevsky et al., 2012]** Krizhevsky, A., Sutskever, I., et Hinton, G. E. (2012). **ImageNet classification with deep convolutional neural networks**. In Pereira, F., Burges, C. J. C., Bottou, L., et Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

**[Lategahn et al., 2013]** Lategahn, H., Beck, J., Kitt, B., et Stiller, C. (2013). **How to learn an illumination robust image feature for place recognition**. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 285–291.

**[Lecun et al., 1998]** Lecun, Y., Bottou, L., Bengio, Y., et Haffner, P. (1998). **Gradient-based learning applied to document recognition**. *Proceedings of the IEEE*, 86(11):2278–2324.

**[Lee et al., 2016]** Lee, E.-K., Gerla, M., Pau, G., Lee, U., et Lim, J.-H. (2016). **Internet of vehicles: From intelligent grid to autonomous cars and vehicular fogs**. *International Journal of Distributed Sensor Networks*, 12(9):1550147716665500.

**[Leutenegger et al., 2011]** Leutenegger, S., Chli, M., et Siegwart, R. Y. (2011). **Brisk: Binary robust invariant scalable keypoints**. In *International Conference on Computer Vision*, pages 2548–2555.

**[Levinson et al., 2011]** Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Kolter, J. Z., Langer, D., Pink, O., Pratt, V., Sokolsky, M., Stanek, G., Stavens, D., Teichman, A., Werling, M., et Thrun, S. (2011). **Towards fully autonomous driving: Systems and algorithms**. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 163–168.

**[Li et al., 2017]** Li, H., Luo, W., et Huang, J. (2017). **Localization of diffusion-based inpainting in digital images**. *IEEE Transactions on Information Forensics and Security*, 12(12):3050–3064.

**[Lin et al., 2017]** Lin, G., Milan, A., Shen, C., et Reid, I. (2017). **Refinenet: Multi-path refinement networks for high-resolution semantic segmentation**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[Lin et al., 2013]** Lin, M., Chen, Q., et Yan, S. (2013). **Network in network**. arXiv:1312.4400.

**[Linegar et al., 2015]** Linegar, C., Churchill, W., et Newman, P. (2015). **Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation**. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 90–97.

**[Liu et al., 2019]** Liu, H., Jiang, B., Xiao, Y., et Yang, C. (2019). **Coherent semantic attention for image inpainting**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

**[Liu et al., 2018]** Liu, J., Yang, S., Fang, Y., et Guo, Z. (2018). **Structure-guided image inpainting using homography transformation**. *IEEE Transactions on Multimedia*, 20(12):3252–3265.

**[Liu et al., 2017]** Liu, M.-Y., Breuel, T., et Kautz, J. (2017). **Unsupervised image-to-image translation networks**. In *Advances in Neural Information Processing Systems 30*, pages 700–708.

**[Liu et al., 2012]** Liu, Y., et Zhang, H. (2012). **Visual loop closure detection with a compact image descriptor**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1051–1056.

**[Long et al., 2015]** Long, J., Shelhamer, E., et Darrell, T. (2015). **Fully convolutional networks for semantic segmentation**. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[Lopez-Antequera et al., 2017]** Lopez-Antequera, M., Gomez-Ojeda, R., Petkov, N., et Gonzalez-Jimenez, J. (2017). **Appearance-invariant place recognition by discriminatively training a convolutional neural network**. *Pattern Recognition Letters*, 92:89–95.

**[Lowe, 1999]** Lowe, D. G. (1999). **Object recognition from local scale-invariant features**. In *Proceedings of the International Conference on Computer Vision (ICCV)*, page 1150, Washington, DC, USA.

**[Lowe, 2004]** Lowe, D. G. (2004). **Distinctive image features from scale-invariant keypoints**. *International Journal of Computer Vision*, 60:91–110.

**[Lowry et al., 2016]** Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., et Milford, M. J. (2016). **Visual place recognition: A survey**. *IEEE Transactions on Robotics*, 32(1):1–19.

**[Luc et al., 2016]** Luc, P., Couprie, C., Chintala, S., et Verbeek, J. (2016). **Semantic segmentation using adversarial networks**. *CoRR*, abs/1611.08408.

**[Maddern et al., 2017]** Maddern, W., Pascoe, G., Linegar, C., et Newman, P. (2017). **1 year, 1000 km: The Oxford RobotCar dataset**. *The International Journal of Robotics Research*, 36(1):3–15.

**[Majdik et al., 2013]** Majdik, A. L., Albers-Schoenberg, Y., et Scaramuzza, D. (2013). **Mav urban localization from google street view data**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3979–3986.

**[Martin et al., 2001]** Martin, D., Fowlkes, C., Tal, D., et Malik, J. (2001). **A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics**. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV*, volume 2, pages 416–423 vol.2.

**[Matas et al., 2004]** Matas, J., Chum, O., Urban, M., et Pajdla, T. (2004). **Robust wide-baseline stereo from maximally stable extremal regions**. *Image and Vision Computing*, 22(10):761–767.

**[McManus et al., 2014]** McManus, C., Upcroft, B., et Newman, P. (2014). **Scene signatures: localised and point-less features for localisation**. In *Robotics: Science and Systems X*, pages 1–9. Robotics: Science and Systems Conference.

**[Menegatti et al., 2004]** Menegatti, E., Maeda, T., et Ishiguro, H. (2004). **Image-based memory for robot navigation using properties of omnidirectional images**. *Robotics and Autonomous Systems*, 47(4):251–267.

**[Merrill et al., 2018]** Merrill, N., et Huang, G. (2018). **Lightweight unsupervised deep loop closure**.

**[Milford et al., 2008]** Milford, M. J., et Wyeth, G. F. (2008). **Mapping a suburb with a single camera using a biologically inspired slam system**. *IEEE Transactions on Robotics*, 24(5):1038–1053.

**[Milford et al., 2012]** Milford, M. J., et Wyeth, G. F. (2012). **Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights**. In *IEEE International Conference on Robotics and Automation*, pages 1643–1649.

**[Milford et al., 2004]** Milford, M. J., Wyeth, G. F., et Prasser, D. (2004). **Ratslam: a hippocampal model for simultaneous localization and mapping**. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04.*, volume 1, pages 403–408 Vol.1.

**[Morel et al., 2009]** Morel, J.-M., et Yu, G. (2009). **Asift: A new framework for fully affine invariant image comparison**. *SIAM Journal on Imaging Sciences*, 2(2):438–469.

**[Naseer et al., 2018]** Naseer, T., Burgard, W., et Stachniss, C. (2018). **Robust visual localization across seasons**. *IEEE Transactions on Robotics*, 34(2):289–302.

**[Nazeri et al., 2019]** Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z., et Ebrahimi, M. (2019). **EdgeConnect: Generative image inpainting with adversarial edge learning**. arXiv:1901.00212.

**[Olid et al., 2018]** Olid, D., Fácil, J. M., et Civera, J. (2018). **Single-view place recognition under seasonal changes**. arXiv:1808.06516.

**[Oliva et al., 2001]** Oliva, A., et Torralba, A. (2001). **Modeling the shape of the scene: A holistic representation of the spatial envelope**. *International Journal of Computer Vision*, 42(3):145–175.

**[Pascoe et al., 2015]** Pascoe, G., Maddern, W., Stewart, A. D., et Newman, P. (2015). **Farlap: Fast robust localisation using appearance priors**. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6366–6373.

**[Pepperell et al., 2014]** Pepperell, E., Corke, P. I., et Milford, M. J. (2014). **All-environment visual place recognition with smart**. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1612–1618.

**[Piasco et al., 2018]** Piasco, N., Sidibé, D., Demonceaux, C., et Gouet-Brunet, V. (2018). **A survey on visual-based localization: On the benefit of heterogeneous data**. *Pattern Recognition*, 74:90–109.

**[Porav et al., 2018]** Porav, H., Maddern, W., et Newman, P. (2018). **Adversarial training for adverse conditions: Robust metric localisation using appearance transfer**. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1011–1018.

**[Prasser et al., 2006]** Prasser, D., Milford, M., et Wyeth, G. (2006). **Outdoor simultaneous localisation and mapping using ratslam**. In *Field and Service Robotics*, pages 143–154, Berlin, Heidelberg. Springer Berlin Heidelberg.

**[Pronobis et al., 2006]** Pronobis, A., Caputo, B., Jensfelt, P., et Christensen, H. I. (2006). **A discriminative approach to robust visual place recognition**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3829–3836.

**[Radenović et al., 2016]** Radenović, F., Tolias, G., et Chum, O. (2016). **Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples**. In *Computer Vision – ECCV 2016*, pages 3–20, Cham. Springer International Publishing.

**[Radenović et al., 2019]** Radenović, F., Tolias, G., et Chum, O. (2019). **Fine-tuning cnn image retrieval with no human annotation**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668.

**[Rajamani, 2006]** Rajamani, R. (2006). **Vehicle Dynamics and Control**.

**[Romero et al., 2010]** Romero, A., et Cazorla, M. (2010). **Topological slam using omnidirectional images: Merging feature detectors and graph-matching**. In *Advanced Concepts for Intelligent Vision Systems*, pages 464–475, Berlin, Heidelberg. Springer Berlin Heidelberg.

**[Ronneberger et al., 2015a]** Ronneberger, O., Fischer, P., et Brox, T. (2015a). **U-net: Convolutional networks for biomedical image segmentation**. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham.

**[Ronneberger et al., 2015b]** Ronneberger, O., Fischer, P., et Brox, T. (2015b). **U-net: Convolutional networks for biomedical image segmentation**. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, pages 234–241, Cham.

**[Rosten et al., 2006]** Rosten, E., et Drummond, T. (2006). **Machine learning for high-speed corner detection**. In *Computer Vision – ECCV 2006*, pages 430–443, Berlin, Heidelberg. Springer Berlin Heidelberg.

**[Rublee et al., 2011]** Rublee, E., Rabaud, V., Konolige, K., et Bradski, G. (2011). **Orb: an efficient alternative to sift or surf**. *International Conference on Computer Vision*, 95:2564–2571.

**[Russakovsky et al., 2015]** Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., et Fei-Fei, L. (2015). **Imagenet large scale visual recognition challenge**. *Int. J. Comput. Vision*, 115(3):211–252.

**[SAE, 2018]** SAE (2018). **Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles**. *SAE Tech. Paper J3016_201806*.

**[Sattler et al., 2017]** Sattler, T., Leibe, B., et Kobbelt, L. (2017). **Efficient & effective prioritized matching for large-scale image-based localization**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1744–1756.

**[Sharif Razavian et al., 2014]** Sharif Razavian, A., Azizpour, H., Sullivan, J., et Carlsson, S. (2014). **Cnn features off-the-shelf: An astounding baseline for recognition**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

**[Shin et al., 2021]** Shin, Y. G., Sagong, M. C., Yeo, Y. J., Kim, S. W., et Ko, S. J. (2021). **Pepsi++: Fast and lightweight network for image inpainting**. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):252–265.

**[Shotton et al., 2009]** Shotton, J., Winn, J., Rother, C., et Criminisi, A. (2009). **Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context**. *International Journal of Computer Vision*, 81(1).

**[Simonyan et al., 2014]** Simonyan, K., et Zisserman, A. (2014). **Very deep convolutional networks for large-scale image recognition**. arXiv:1409.1556.

**[Singh, 2015]** Singh, S. (2015). **Critical reasons for crashes investigated in the national motor vehicle crash causation survey**. *Washington, DC, USA*, Tech. Rep. DOT HS 812 115.

**[Sivic et al., 2003]** Sivic, et Zisserman (2003). **Video google: a text retrieval approach to object matching in videos**. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2.

**[Svärm et al., 2017]** Svärm, L., Enqvist, O., Kahl, F., et Oskarsson, M. (2017). **City-scale localization for cameras with known vertical direction**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1455–1461.

**[Szegedy et al., 2015]** Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., et Rabinovich, A. (2015). **Going deeper with convolutions**. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[Sünderhauf et al., 2011]** Sünderhauf, N., et Protzel, P. (2011). **Brief-gist - closing the loop by simple means**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1234–1241.

**[Sünderhauf et al., 2015a]** Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., et Milford, M. (2015a). **On the performance of ConvNet features for place recognition**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4297–4304.

**[Sünderhauf et al., 2015b]** Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., et Milford, M. (2015b). **Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free**. In *Robotics: Science and Systems XI*, pages 1–10. Robotics: Science and Systems Conference.

**[Tao et al., 2018]** Tao, X., Gao, H., Shen, X., Wang, J., et Jia, J. (2018). **Scale-recurrent network for deep image deblurring**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[Torii et al., 2015]** Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., et Pajdla, T. (2015). **24/7 place recognition by view synthesis**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[Trzcinski et al., 2012]** Trzcinski, T., et Lepetit, V. (2012). **Efficient discriminative projections for compact binary descriptors**. In *Computer Vision – ECCV 2012*, pages 228–242, Berlin, Heidelberg. Springer Berlin Heidelberg.

**[Ulmer, 1994]** Ulmer, B. (1994). **Vita ii-active collision avoidance in real traffic**. In *Proceedings of the Intelligent Vehicles '94 Symposium*, pages 1–6.

**[Ulrich et al., 2000]** Ulrich, I., et Nourbakhsh, I. (2000). **Appearance-based place recognition for topological localization**. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, volume 2, pages 1023–1029 vol.2.

**[Urmson et al., 2004]** Urmson, C., Anhalt, J., Clark, M., Galatali, T., Gonzalez, J. P., Gowdy, J., Gutierrez, A., Harbaugh, S., Johnson-Roberson, M., Kato, H., Koon, P. L., Peterson, K., Smith, B. K., Spiker, S., Tryzelaar, E., et Whittaker, W. R. L. (2004). **High speed navigation of unrehearsed terrain: Red team technology for grand challenge 2004**. Technical Report CMU-RI-TR-04-37, Carnegie Mellon University.

**[Valgren et al., 2007]** Valgren, C., et Lilienthal, A. J. (2007). **Sift, surf and seasons : long-term outdoor localization using local features**. In *ECMR 2007 : Proceedings of the European Conference on Mobile Robots*, pages 253–258.

**[Wang et al., 2018]** Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., et Catanzaro, B. (2018). **High-resolution image synthesis and semantic manipulation with condi-**

**tional GANs**. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[Warren et al., 2010]** Warren, M., McKinnon, D., He, H., et Upcroft, B. (2010). **Unaided stereo vision based pose estimation**. In Wyeth, G., et Upcroft, B., editors, *Australasian Conference on Robotics and Automation*, Brisbane. Australian Robotics and Automation Association.

**[Weerasekera et al., 2018]** Weerasekera, C. S., Dharmasiri, T., Garg, R., Drummond, T., et Reid, I. (2018). **Just-in-time reconstruction: Inpainting sparse maps using single view depth predictors as priors**. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4977–4984.

**[Wei et al., 2013]** Wei, J., Snider, J. M., Kim, J., Dolan, J. M., Rajkumar, R., et Litkouhi, B. (2013). **Towards a viable autonomous driving research platform**. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 763–770.

**[Wu et al., 2019]** Wu, L., et Wu, Y. (2019). **Deep supervised hashing with similar hierarchy for place recognition**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3781–3786.

**[Xin et al., 2019]** Xin, Z., Cai, Y., Lu, T., Xing, X., Cai, S., Zhang, J., Yang, Y., et Wang, Y. (2019). **Localizing discriminative visual landmarks for place recognition**. In *International Conference on Robotics and Automation (ICRA)*, pages 5979–5985.

**[Xiong et al., 2019]** Xiong, W., Yu, J., Lin, Z., Yang, J., Lu, X., Barnes, C., et Luo, J. (2019). **Foreground-aware image inpainting**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[Xu et al., 2014]** Xu, D., Badino, H., et Huber, D. (2014). **Topometric localization on a road network**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3448–3455.

**[Xue et al., 2017a]** Xue, H., Zhang, S., et Cai, D. (2017a). **Depth image inpainting: Improving low rank matrix completion with low gradient regularization**. *IEEE Transactions on Image Processing*, 26(9):4311–4320.

**[Xue et al., 2017b]** Xue, H., Zhang, S., et Cai, D. (2017b). **Depth image inpainting: Improving low rank matrix completion with low gradient regularization**. *IEEE Transactions on Image Processing*, 26(9):4311–4320.

**[Ying et al., 2017]** Ying, Z., Li, G., et Gao, W. (2017). **A bio-inspired multi-exposure fusion framework for low-light image enhancement**. arXiv:1711.00591.

**[Yurtsever et al., 2020]** Yurtsever, E., Lambert, J., Carballo, A., et Takeda, K. (2020). **A survey of autonomous driving: Common practices and emerging technologies**. *IEEE Access*, 8:58443–58469.

**[Zeng et al., 2019]** Zeng, Y., Fu, J., Chao, H., et Guo, B. (2019). **Learning pyramid-context encoder network for high-quality image inpainting**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

**[Zhang et al., 2020]** Zhang, A., Lipton, Z. C., Li, M., et Smola, A. J. (2020). **Dive into Deep Learning**. http://www.d2l.ai.

**[Zhang et al., 2012]** Zhang, F., Stähle, H., Chen, G., Simon, C. C. C., Buckl, C., et Knoll, A. (2012). **A sensor fusion approach for localization with cumulative error elimination**. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 1–6.

**[Zhang et al., 2017]** Zhang, X., Su, Y., et Zhu, X. (2017). **Loop closure detection for visual slam systems using convolutional neural network**. In *23rd International Conference on Automation and Computing (ICAC)*, pages 1–6.

**[Zhang et al., 2021]** Zhang, X., Wang, L., et Su, Y. (2021). **Visual place recognition: A survey from deep learning perspective**. *Pattern Recognition*, 113:107760.

**[Zhao et al., 2019]** Zhao, C., Ding, R., et Key, H. L. (2019). **End-to-end visual place recognition based on deep metric learning and self-adaptively enhanced similarity metric**. In *IEEE International Conference on Image Processing (ICIP)*, pages 275–279.

**[Zhao et al., 2017]** Zhao, H., Shi, J., Qi, X., Wang, X., et Jia, J. (2017). **Pyramid scene parsing network**. In *CVPR*.

**[Zhao et al., 2018]** Zhao, J., Chen, Z., Zhang, L., et Jin, X. (2018). **Unsupervised learnable sinogram inpainting network (sin) for limited angle ct reconstruction**.

**[Zhou et al., 2018]** Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., et Torralba, A. (2018). **Places: A 10 million image database for scene recognition**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6).

**[Zhu et al., 2017]** Zhu, J.-Y., Park, T., Isola, P., et Efros, A. A. (2017). **Unpaired image-to-image translation using cycle-consistent adversarial networks**. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

**[Zhu et al., 2018]** Zhu, X., Qian, Y., Zhao, X., Sun, B., et Sun, Y. (2018). **A deep learning approach to patch-based image inpainting forensics**. *Signal Processing: Image Communication*, 67:90 – 99.

# List of Figures

# List of Tables

# A

# PUBLICATIONS

## A.1/ CONFERENCES

L. Liu, C. Cappelle and Y. Ruichek, **Day and Night Place Recognition Based on Low-quality Night-time Images**. IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), 2020, pp. 1-6.

L. Liu, C. Cappelle and Y. Ruichek, **Image-Based Place Recognition Using Semantic Segmentation and Inpainting to Remove Dynamic Objects**. Image and Signal Processing, Springer International Publishing, 2020, pp. 262–270.

## A.2/ JOUNALS

L. Liu, C. Cappelle, and Y. Ruichek, **Improving place recognition by removing dynamic objects using inpainted images**. Journal of Visual Communication and Image Representation (JVCIR), under review.

Document generated with LaTeX and:
the LaTeX style for PhD Thesis created by S. Galland — http://www.multiagent.fr/ThesisStyle
the tex-upmethodology package suite — http://www.arakhne.org/tex-upmethodology/