More Robust Visual Place Recognition with Image-to-Image Augmentations from Vision Foundation Models

by

Fabian Gebben

Student number:	5174198
Programme:	MSc Robotics
Faculty:	Faculty of Mechanical Engineering, TU Delft
Department:	Cognitive Robotics
Defense date:	June 6th 2025
Graduation committee:	Dr. ir. J.F.P. Kooij (TU Delft) Ir. M. Zaffar (TU Delft) Dr. ir. S. Khademi (TU Delft)

More Robust Visual Place Recognition with Image-to-Image Augmentations from Vision Foundation Models

Fabian Gebben (5174198)
Supervisor: Dr. ir. J.F.P. Kooij, TU Delft ME
Daily supervisor: Ir. M. Zaffar, TU Delft ME
Delft University of Technology - Department of Cognitive Robotics

***Abstract*— Visual Place Recognition (VPR) remains a challenging problem, particularly under difficult conditions such as night-time or winter weather, which are often underrepresented in existing training datasets. Although transformer-based models have recently advanced the state-of-the-art, their high computational demands can hinder deployment in real-world robotic systems. This thesis proposes a new data augmentation strategy for VPR using image-to-image Vision Foundation Model InstructPix2Pix to generate realistic visual variations such as night and snow scenes from the original training data. These synthetic augmentations are added to the original training dataset to extend dataset diversity without requiring additional data collection. To further improve performance, the method is combined with more advanced augmentations using the Kornia library, which already improves robustness over traditional augmentation techniques. Experiments on multiple benchmark datasets show that lightweight, ResNet-based models trained with our VFM augmentations achieve significantly improved performance under challenging visual conditions. Additional ablations demonstrate the importance of careful prompt design and hyperparameter tuning. Overall, this work shows that VFMs can serve as practical tools for targeted dataset augmentation, improving the robustness of existing VPR methods in difficult scenarios.**

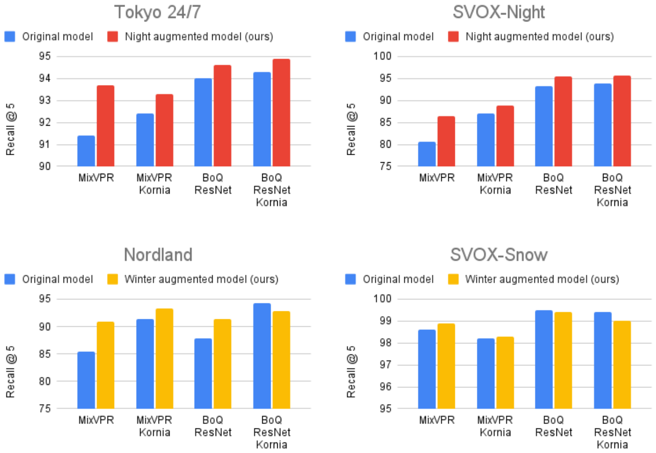

Fig. 1: Recall@5 performance across four benchmark datasets [4, 5, 6], comparing original models to those trained with our VFM-based image-to-image augmentation strategy (red for night-time augmentation, yellow for winter). The results show that incorporating VFM-augmented training data consistently improves performance for ResNet-based VPR models MixVPR [7] and BoQ ResNet [8], under challenging night-time and winter conditions.

I. Introduction

The ability to recognize a specific location solely from visual information in an image, known as Visual Place Recognition (VPR), is a fundamental component of robot state estimation [1]. A VPR system attempts to identify the location of a query image by matching it to the most similar image in a large reference database. VPR plays a critical role in a range of applications, from autonomous navigation to augmented reality [2]. The most common application is in Simultaneous Localization and Mapping (SLAM) [3], where VPR is used for loop closure detection to reduce localization drift in robotic navigation systems [1].

In many real-world robot applications, where robots and embedded platforms have limited onboard processing power and must make quick localization decisions, efficiency is crucial. Because of this, memory usage and inference speed are often just as important as accuracy. VPR methods deployed in these settings must therefore find a balance between accuracy and computational efficiency [9, 10].

Early VPR techniques relied on hand-crafted feature descriptors to extract local and global image features [1]. However, these approaches were limited in their ability to capture high-level semantic information and struggled to adapt to the specific challenges of VPR due to their general-purpose design [2]. The field has since shifted toward deep learning-based approaches, particularly those using Convolutional Neural Networks (CNNs) [11, 12], which extract more expressive visual features and can be fine-tuned for the VPR-specific task. CNN-based models and especially those based on the ResNet [13] architecture have demonstrated improved robustness and adaptability, while remaining lightweight enough for real-time deployment [2].

Recently, Vision Transformers (ViTs) have emerged as a powerful alternative to CNNs for feature extraction in VPR. Unlike CNNs, which are restricted by local receptive fields, ViTs leverage self-attention mechanisms to capture global relationships across the entire image [14], which allows them to significantly outperform CNN-based VPR methods [8, 15].

However, these ViT-based models typically require significantly more computational resources, both in terms of memory and inference time [16, 17]. This makes them less suitable for certain real-time robotic applications, where lightweight CNN-based models can remain more practical.

Despite advancements in model architecture, modern VPR systems still struggle under challenging conditions. Performance can decrease significantly due to factors such as viewpoint changes, lighting variations, seasonal transitions, weather conditions, dynamic objects, and occlusions [1]. Although several VPR datasets exist, most fail to represent all these challenges typically encountered in VPR. In particular, extreme lighting conditions, unusual viewpoints, and drastic seasonal shifts are often underrepresented in the most popular training datasets, which limits the ability of models to learn and generalize to these scenarios [18]. This limitation is especially critical for ResNet-based models, which are pre-trained on ImageNet [19], a dataset that also lacks many of these challenges. When both pre-training and fine-tuning datasets fail to expose models to scenarios like nighttime or winter environments, reliable performance under such conditions becomes unlikely during deployment.

To address these limitations, we propose extending existing VPR training datasets by using an image-to-image Vision Foundation Model (VFM). By leveraging the generative capabilities of VFMs, we create new data that simulates challenging conditions such as day-to-night transitions or seasonal changes. Unlike conventional data augmentation, this approach introduces genuinely new visual content to the original image while preserving the scene's spatial layout and semantics. This enables the controlled introduction of hard-to-capture conditions into existing training datasets.

Our goal is to evaluate whether this form of targeted, VFM-based augmentation can improve the performance of more lightweight, ResNet-based VPR models, helping them better handle certain VPR challenges without increasing computational cost and closing the performance gap to more computationally expensive transformer-based methods, which leads to the following research question:

> *"How can image-to-image Vision Foundation Models be used for dataset augmentation to improve ResNet-based Visual Place Recognition methods?"*

The remainder of this paper is structured as follows: Section II reviews relevant related work. Section III then presents the methodology in detail. Section IV describes the experimental setup and Section V reports the results. Finally, Section VI concludes the paper and outlines future research directions.

II. Related work

A. Traditional methods

The first VPR methods primarily relied on handcrafted feature descriptors to extract visual information from images. Local feature extraction techniques, such as SIFT [20] and SURF [21], were commonly used to detect keypoints and compute local descriptors for these keypoints. Other traditional approaches focused on global descriptors, which aimed to summarize the visual content of an entire image. Methods like Bag-of-Words (BoW) [22], VLAD [23], and GIST [24] created compact global feature representations that could be used for image matching. While handcrafted methods performed reasonably well under stable conditions, they struggled with the variability of real-world environments. Their general-purpose design limited their ability to handle typical VPR challenges such as viewpoint changes and lighting variations, which motivated the transition to more robust deep-learning-based approaches [25].

B. CNN-based methods

The rise of deep learning introduced a significant shift in VPR, with Convolutional Neural Networks becoming very popular for feature extraction in VPR models. CNNs automatically learn hierarchical feature representations from raw image data, effectively capturing complex visual patterns essential for VPR tasks [25]. Early VPR approaches which used CNNs often employed pre-trained networks, such as VGG [26] and ResNet [13], as fixed feature extractors [27]. Newer methods improved on this by fine-tuning these pre-trained backbones on VPR-specific datasets, often integrating a specialized aggregation technique to create more robust data-driven models [12, 28].

As deep learning techniques evolved, more sophisticated CNN-based models emerged, such as MixVPR [7], EigenPlaces [29] and BoQ [8], each contributing to enhanced robustness and accuracy in VPR tasks. However, all these CNN-based methods have the same limitation: they rely on ResNet-based backbone networks, which are pre-trained on the ImageNet dataset, which lacks most of the diverse environmental conditions typically encountered in VPR problems [19]. This means that these models are not exposed to challenges like night-time scenes or seasonal changes during pre-training. If such conditions are also missing from the dataset used for fine-tuning, the models struggle to perform well in these scenarios, leading to suboptimal performance in real-world applications [30].

C. Transformer based methods

The introduction of transformer architectures, particularly Vision Transformers (ViTs), has marked a significant advancement in the field of VPR. Unlike CNNs, which are constrained by local receptive fields, transformers utilize self-attention mechanisms to capture global dependencies across entire images, enabling more comprehensive feature representations [14].

The first application of transformers in VPR was AnyLoc [15], which leverages pre-trained transformer models such as DINOv2 [31], CLIP [32] and MAE [33] to extract features without additional fine-tuning. This approach has demonstrated robust performance across diverse environments and conditions, highlighting the potential of transformer backbones in VPR applications, even without fine-tuning.

More recent methods like SALAD [34], CricaVPR [35] and DINOv2 BOQ [8] have fine-tuned these transformer backbones on VPR-specific datasets. These approaches have achieved new state-of-the-art performance, significantly outperforming earlier CNN-based methods.

However, the enhanced capabilities of transformer-based models come with increased computational demands. Transformers typically require more memory and exhibit slower inference times compared to CNNs, posing challenges for deployment in resource-constrained environments such as real-time robotic systems [16, 17].

D. Data augmentation and dataset extension

Data augmentation is a widely used technique in deep learning to improve model generalization and robustness, especially when data collection is expensive or limited [36] and is also a standard component of most VPR pipelines. Techniques such as random cropping, flipping, color jittering, and brightness adjustment are often applied as part of the training process [37]. However, while augmentation is widely used in VPR, the majority of these methods remain relatively traditional and simplistic, focusing on low-level transformations rather than introducing realistic, high-level scene variations.

Some studies have explored more systematic or advanced augmentation techniques. For example, one study [37] evaluated geometric transformations, illumination changes, and occlusion methods such as Cutout [38] and GridMask [39], showing modest improvements in performance, especially when combined with techniques like RandAugment. Others have looked into query-specific augmentations like color jittering and flipping [40], or style-based methods such as style transfer [5] and style randomization [41], which aim to improve domain generalization.

An example of a more advanced augmentation in VPR is CLASP-Net [42], which uses the Kornia library [43]. Kornia is a toolkit that enables the application of a broader, more advanced and more diverse set of pixel-level transformations, helping the model learn robustness to appearance variations. Although their approach demonstrated the potential of using Kornia augmentations for improved VPR performance, the use of Kornia remains underexplored in other VPR methods, as most VPR pipelines continue to rely on simpler augmentation methods.

Despite these advancements, current augmentation techniques used in VPR are not able to simulate very realistic, semantically rich conditions such as night-time scenes, seasonal shifts, or adverse weather. As a result, there remains a gap between what current augmentation methods can offer and the kind of complex, photorealistic variability encountered in real-world VPR applications.

E. Image-to-image Vision Foundation Models

The emergence of Vision Foundation Models (VFMs) has significantly expanded the possibilities for image generation and editing tasks in computer vision. Trained on large-scale, diverse datasets, these models are designed to generalize across a wide range of visual domains and downstream tasks [44]. Of particular interest in the context of this thesis are image-to-image VFMs, which take an existing image and a text prompt as input and produce a semantically modified version of the image as output.

Unlike traditional data augmentation methods, which rely on geometric transformations or simple color and brightness changes, image-to-image VFMs can introduce high-level, photorealistic transformations such as turning day into night, adding snow, or changing architectural styles. This capability could be highly beneficial for VPR, where generalization across diverse environmental conditions is essential, and where many datasets are missing some of these challenges.

Several image-to-image models have recently been developed, with most methods building on top of the open source VFM of Stable Diffusion [45]. For instance, DA-Fusion [46] uses Stable Diffusion to generate diverse semantic variations of input images, improving generalization in low-data regimes. InstructPix2Pix [47] combines Stable Diffusion with GPT-3 [48] to allow natural language-based edits, providing fine control over how images are transformed while preserving spatial consistency. Imagic [49] introduces more precise, text-guided edits through optimized text embeddings, enabling subtle yet semantically meaningful changes. Finally, StreamDiffusion [50] focuses on real-time image generation, prioritizing speed and efficiency through batched denoising strategies, though possibly at some cost to image quality.

While existing VPR state-of-the-art methods often rely on traditional augmentation techniques and high-capacity transformer-based models to improve robustness, this thesis explores an alternative strategy: using image-to-image vision foundation models to generate realistic, semantically meaningful augmentations from existing images. This approach aims to enrich training datasets with underrepresented conditions, such as night-time and snow, without the need for costly data collection or complex architectures.

By integrating these VFM-generated augmentations into the training pipeline, we focus primarily on improving the performance of more lightweight, ResNet-based VPR models, showing that when trained on richer, more diverse data, these models can close the gap with state-of-the-art transformer-based methods.

In this work, we present the following major contributions:

- A novel data augmentation strategy for VPR using image-to-image vision foundation models to generate realistic variations (e.g., night-time, winter) is introduced, which can significantly improve performance on ResNet-based models, narrowing the gap with transformer-based methods.
- A systematic ablation study of prompt design and key hyperparameters (e.g., text/image weights, diffusion steps, augmentation ratio) is conducted to evaluate and optimize the use of image-to-image vision foundation models for data augmentation in VPR.
- The effectiveness of replacing traditional augmentations with Kornia-based augmentations is demonstrated, showing that Kornia offers a superior alternative for enhancing the robustness of existing ResNet-based VPR methods.

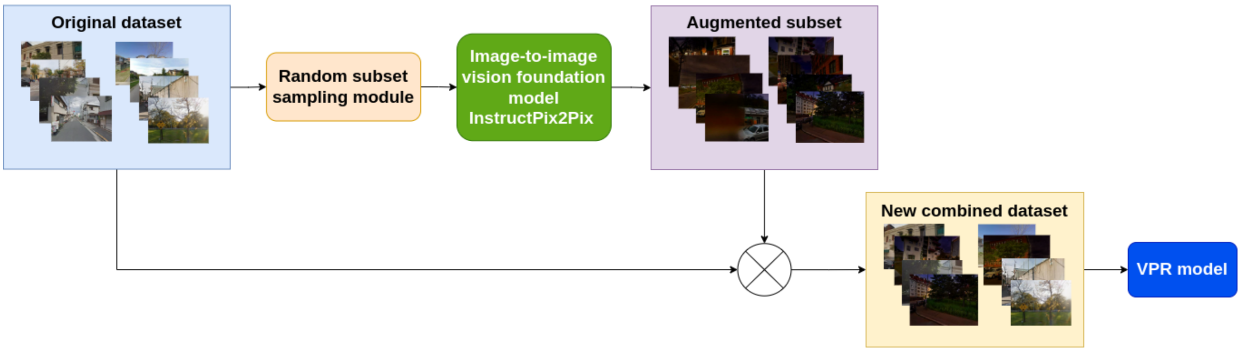
Original dataset
Random subset sampling module
Image-to-image vision foundation model InstructPix2Pix
Augmented subset
New combined dataset
VPR model

Fig. 2: Overview of our pipeline used to introduce new challenges into existing datasets by using image-to-image vision foundation model InstructPix2Pix [47] to create partially augmented training datasets.

III. Methodology

A. Dataset selection

To effectively evaluate the impact of image-to-image VFM-based augmentation, it is important to select a training dataset that can benefit from the inclusion of certain challenging conditions. In this study, we focus on night-time and snow, which are commonly underrepresented in existing VPR training datasets [18]. These conditions will be synthetically introduced into the training data using a VFM-based approach.

To assess whether this augmentation leads to improved model performance, we also identify appropriate benchmark datasets that include real-world instances of these challenges. These test datasets enable a direct evaluation of the effectiveness of the augmented training data.

The specific training and evaluation datasets, along with the motivation for their selection, are discussed in detail in Chapter IV-A.

B. Image-to-image model selection and parameter selection

To introduce missing visual challenges into the training dataset, image-to-image VFM InstructPix2Pix [47], a powerful generative model that builds on Stable Diffusion [45] and GPT-3 [48], is used. InstructPix2Pix enables high-quality image transformations guided by natural language instructions, while preserving the semantic and structural content of the original image. The model is open-source and relatively efficient, allowing for fast generation of synthetic images at scale. These properties make it particularly well-suited for augmenting VPR datasets with realistic scene variations, such as night-time or winter weather, that are typically underrepresented in existing training data [18].

InstructPix2Pix operates by taking an input image along with a textual prompt that describes the desired transformation. It then generates a modified image that reflects the requested change while maintaining the spatial layout and key semantic features of the scene. To ensure that the generated images were both realistic and relevant for VPR tasks, a hyperparameter tuning study was conducted (details provided in Section V-C) to find a good balance between keeping the relevant semantic information and adding the challenge to the image.

C. Dataset extension and training on extended datasets

Once the vision foundation model and hyperparameters were selected, a data augmentation pipeline was implemented to integrate synthetic images into the training process as is shown in Figure 2. The objective was to enrich the dataset with realistic environmental variations, specifically night-time and winter scenes, while preserving spatial consistency and ground-truth labels.

The process begins by randomly selecting approximately one-sixth of the images from the original training dataset as candidates for augmentation. Each selected image is then processed using the InstructPix2Pix model, guided by predefined prompts and tuned parameters, to simulate either a night-time or snow-covered version, while trying to retain the spatial structure and scene identity of the original image.

Rather than replacing existing data, the synthetic images are added on top of the original training set. This preserves the full diversity of the dataset while introducing new realistic examples of underrepresented conditions. The final training set consists of the complete original dataset, supplemented with transformed samples, such that approximately one in every seven images presents a synthetically generated visual challenge. This augmentation strategy exposes the model to a broader range of visual appearances, with the goal of improving generalization to unseen environments, particularly those involving lighting and seasonal variation common in real-world deployments without requiring any changes to the VPR model architecture. As a result, it can be seamlessly combined with traditional data augmentation techniques or more advanced augmentation techniques such as Kornia [43].

IV. Experimental setup

A. Datasets

To evaluate the impact of incorporating synthetically augmented images into existing datasets during training, a set of suitable VPR datasets was selected. Most importantly, a training dataset was required in which specific challenges such as night-time conditions and seasonal changes were missing or underrepresented.

It was also important to choose relevant evaluation datasets in which these same challenges do exist. This will enable enable a clear comparison: if training on the partially augmented

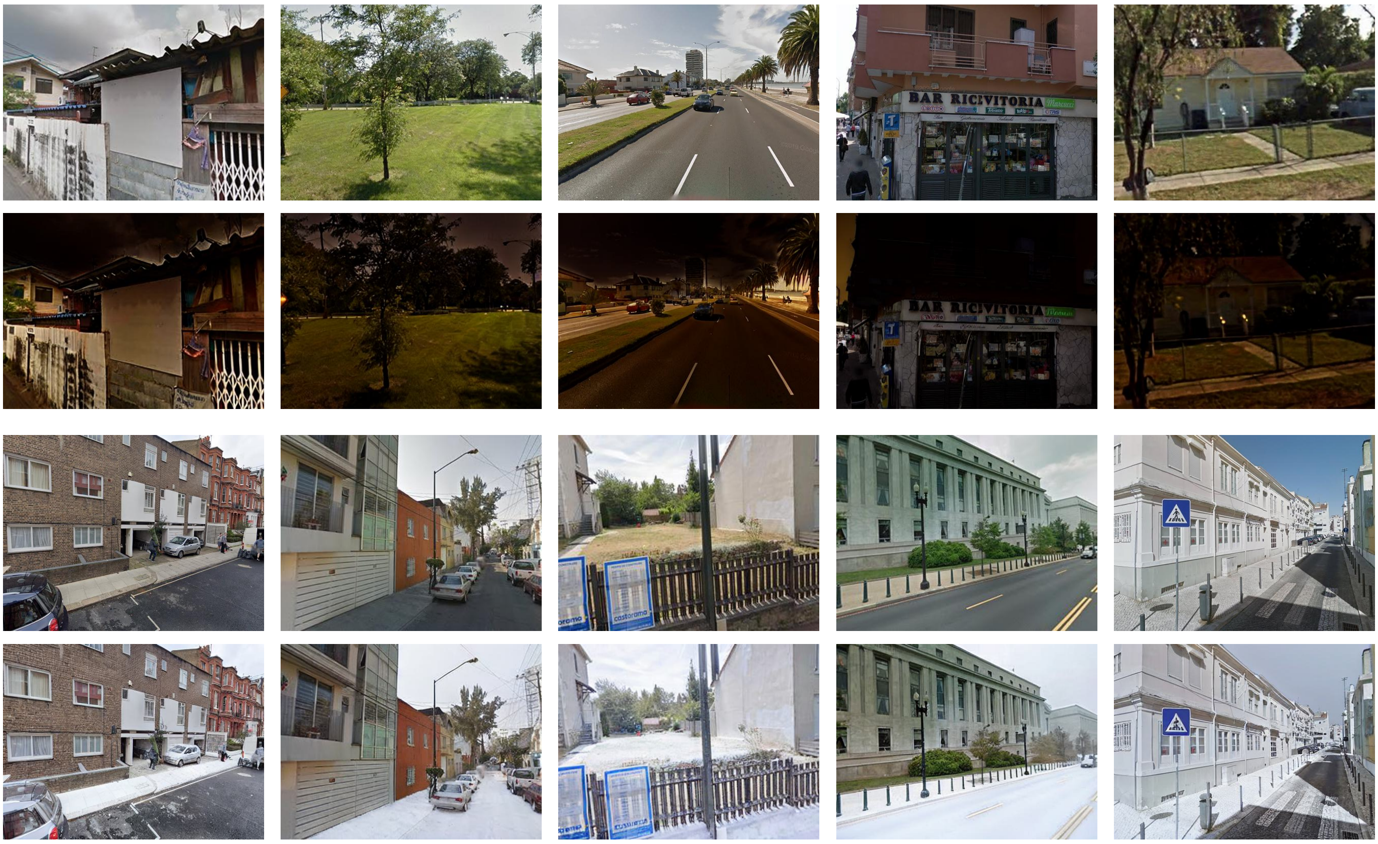

Fig. 3: Examples of the original GSV-cities images converted to night (first two rows) or winter (bottom two rows) using InstructPix2Pix [47].

dataset is effective, models trained on it should perform better on these relevant test sets than those that are only trained on the original dataset.

The following sections describe the datasets used in this thesis, their roles in training and evaluation, and the rationale behind their selection with an overview given in Table I.

1) GSV-cities: GSV-cities [51] is a large-scale training dataset designed for VPR research. It consists of approximately 530,000 Google Street View images collected from over 40 cities worldwide, representing more than 62,000 distinct locations. Each location is depicted by multiple images captured over a 14-year span with highly accurate ground truth data.

This dataset is the most popular dataset used for training VPR methods [7, 8, 35, 51] due to its extensive geographic coverage and substantial size, which strikes a balance between diversity and manageability. However, since it is derived from Google Street View, the dataset lacks images captured during nighttime or winter conditions. These scenarios present significant challenges for VPR systems in real-world deployments.

This gap presents an opportunity to augment the GSV-Cities dataset by artificially introducing these missing challenges. By employing vision foundation models, we can generate synthetic nighttime and winter images from the existing data. This augmentation aims to enhance the dataset's diversity, thereby improving the robustness and generalization capabilities of VPR models trained on it.

2) Tokyo 24/7: Tokyo 24/7 [4] is a dataset specifically designed to evaluate VPR systems under varying illumination conditions, including significant day-to-night changes. The dataset comprises 315 query images captured at 125 distinct locations throughout Tokyo. At each location, images were taken from three different viewpoints at three different times of day: daytime, sunset, and night. This setup provides a comprehensive assessment of temporal variations in urban environments.

The reference database consists of 76k geo-tagged images collected from Google Street View, offering a broad representation of the city's scenes under standard conditions. The combination of diverse viewpoints and illumination conditions makes Tokyo 24/7 a challenging and valuable benchmark for assessing the robustness of VPR systems, particularly in handling day-to-night transitions.

In this study, we use Tokyo 24/7 to evaluate our model's performance in recognizing places under day-to-night changes.

3) SVOX-Night and SVOX-Snow: SVOX (Street View Oxford) [5] is a test dataset specifically designed for evaluating cross-domain VPR. It is built by combining reference images from Google Street View, covering the city of Oxford, with query images from the Oxford RobotCar dataset [52]. While GSV-Cities includes images from many cities worldwide, it does not include Oxford, ensuring that there is no data overlap between the training and evaluation datasets in this study.

The RobotCar queries are labeled with their lighting and weather conditions (Snow, Rain, Sun, Night or Overcast). This labeling enables targeted evaluation under specific visual challenges. In this work, we focus specifically on the night-time and snow conditions, as these represent some of the most difficult scenarios for VPR systems.

For the experiments in this thesis, we use the 823 night-time queries and 870 snow queries from RobotCar in the SVOX test set and attempt to match each to its corresponding location in the SVOX gallery, which contains 17,166 reference images taken from Google Street View.

4) Nordland: Nordland [6] captures a train journey of 728 km through different seasons and provides around 115k images. For evaluation, the dataset is scaled-down to a version with 27k reference images and 27k query images. The reference images are taken in the summer and the query images are taken in the winter [53]. The challenge in this dataset is thus to match an image taken in snowing winter conditions to a summer image. This means that this dataset can be used to show how well the model can handle this domain-gap between winter and summer images.

Dataset	**# Query images**	**# Database images**	**Night**	**Winter**
GSV-cities [51]	529683			
Tokyo 24/7 [4]	315	75984	✓	
SVOX-Night [5]	823	17166	✓	
Nordland [6]	27592	27592		✓
SVOX-Snow [5]	870	17166		✓

TABLE I: Overview of datasets used for training and evaluation.

B. Evaluated Techniques

1) MixVPR: MixVPR [7] is one of the most advanced approaches based on a ResNet CNN backbone. It introduces a novel feature aggregation mechanism using multiple feature mixing blocks, which capture global relationships between features extracted from pre-trained backbones. Combined with a Multi-Similarity loss function [54] and trained on the GSV-Cities dataset [51], MixVPR achieves competitive performance while remaining relatively lightweight.

Its ability to mix features across different abstraction levels enables strong generalization, even under limited computational resources. However, its performance has started to lag behind newer transformer-based models, also on night and winter datasets, making it a strong candidate for evaluating the effectiveness of our augmentation method.

2) BoQ: BoQ (Bag of Learnable Queries) [8] also used a ResNet-based architecture and introduces a more advanced aggregation technique via a multi-head attention (MHA) mechanism [55]. The model incorporates learnable global queries that interact with feature maps through cross-attention, enabling the selective aggregation of the most relevant image features. The output is a compact, discriminative global descriptor formed through concatenation, projection, and normalization. Like MixVPR, it is trained using the Multi-Similarity loss and on the GSV-Cities dataset.

BoQ currently represents the state-of-the-art for ResNet-based VPR methods and has demonstrated strong results across diverse datasets. Its attention-based design helps it handle complex scene structures more effectively. However, challenges like night-time and winter conditions still impact its robustness, making it another suitable benchmark for evaluating the impact of our data augmentation strategy.

C. Implementation details

To generate realistic synthetic images that effectively introduce the new visual challenges, a small hyperparameter study was conducted on the most optimal use of the InstructPix2Pix model [47]. The results of this tuning are discussed in more detail in Section V-C. Three core hyperparameters were varied: the image guidance scale, which controls how closely the output resembles the original image; the text guidance scale, which determines the strength of adherence to the textual prompt; and the number of diffusion steps, which influences the visual quality and transformation extent. In addition to these parameters, careful prompt design played a critical role in achieving realistic augmentations.

Table II summarizes the selected prompts and corresponding hyperparameters used to generate night-time and winter versions of original database images.

Type of Augmentation	**Prompt**	**Image Weight**	**Text Weight**	**Diffusion Steps**
Night	"It is now midnight"	1.5	15.0	20
Winter	"It is now snowing"	1.2	15.0	20

TABLE II: Hyperparameters and prompts used to generate augmented night and winter images from original GSV-cities dataset images using InstructPix2Pix [47].

The GSV-Cities dataset is extended by randomly sampling one in six images from the original dataset and transforming them into either night-time or winter-themed images using the prompts and hyperparameters listed in Table II. Example outputs of these transformations are shown in Figure 3. The augmented images are added on top of the original dataset, increasing the total dataset size by approximately 16.67%. This means that around one in seven training images is a synthetically augmented image generated by InstructPix2Pix [47].

Unlike many standard VPR datasets, GSV-Cities does not provide a fixed query/reference split. Instead, the splits are randomly redefined at the beginning of each epoch. As a result, the augmented images can serve as either query or reference images throughout training.

Both MixVPR and BoQ models are trained using the original GSV-Cities dataloaders provided by the respective authors. For MixVPR, training follows the original setup exactly. A ResNet50 backbone is used to generate 4096-dimensional descriptors, and training is performed for 80 epochs. The model is validated on the Pitts30k dataset [56], and the checkpoint with the highest Recall@1 score on this validation set is selected for testing.

BoQ is similarly trained with a ResNet50 backbone and largely follows the configuration used by the original authors, with one exception: the batch size is reduced from 128 to 64 to fit within the available GPU memory. The descriptor size

Model	Augmentations Used			Night datasets				Winter datasets				Average performance	
	Traditional	Kornia	VFM (ours)	SVOX-Night		Tokyo 24/7		Nordland		SVOX-Snow			
				R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
MixVPR	✓			63.2	80.7	81.6	91.4	73.3	85.5	**96.6**	98.6	78.7	89.1
	✓		✓	74.4	86.5	85.1	**93.7**	81.8	90.9	96.4	**98.9**	84.4	92.5
		✓		76.4	87.1	**87.0**	92.4	83.3	91.3	94.1	98.2	85.2	92.2
		✓	✓	**80.0**	**88.8**	86.0	93.3	**86.4**	**93.3**	94.4	98.3	**86.7**	**93.4**
ResNet BoQ	✓			85.5	93.2	**89.9**	94.0	78.4	87.9	**98.5**	**99.5**	88.1	93.7
	✓		✓	89.4	95.4	89.2	94.6	83.3	91.3	**98.5**	99.4	90.1	95.2
		✓		88.6	93.8	88.3	94.3	85.3	**94.3**	97.8	99.4	90.0	95.4
		✓	✓	**91.1**	**95.6**	88.6	**94.9**	**86.8**	92.8	97.7	99.0	**91.0**	**95.6**

TABLE III: Comparison of Recall@N performance between MixVPR and ResNet BoQ models under various augmentation strategies, including traditional augmentations, Kornia [43], and our proposed Vision Foundation Model (VFM) augmentations. For VFM augmentations, a night-augmented model is used for evaluation on night datasets, and a winter-augmented model is used for evaluation on winter datasets. Evaluation is conducted on four benchmark datasets. Best results are shown in **bold**, and second-best results are underlined.

is set to its maximum of 16384. Training also runs for 80 epochs, and the best model is selected based on Recall@1 performance on the MSLS-val dataset [18].

To evaluate the impact of our VFM-based augmentation method, we compare model performance across four different training conditions: (1) using only the original traditional augmentations, (2) using the traditional augmentations combined with our proposed VFM-based augmentations, (3) using Kornia-based augmentations [43] instead of the traditional augmentations and (4) using Kornia-based augmentations combined with our VFM-based augmentations. For the traditional augmentation baseline, we retain the RandAugment configuration used in the original MixVPR and BoQ implementations. Kornia augmentations are selected based on those used in prior work by [42].

All models are trained and evaluated on an NVIDIA Tesla V100-SXM2-32GB GPU. Model performance on test datasets is evaluated using the standard Recall@k metric, as implemented in the VPR evaluation framework used for evaluation [29]. Recall@k reports the percentage of query images for which at least one of the top-k retrieved reference images is within a ground truth threshold of 25 meters [1]. For the Nordland dataset, a query is considered correct if the reference lies within 10 frames of the query image [6].

V. Results and Analysis

A. *Quantitative results*

Table III reports Recall@1 and Recall@5 scores for both MixVPR and BoQ on four datasets that capture the challenging night-time and winter conditions. Figure 1 also provides a visual comparison of Recall@5 performance when using traditional or Kornia augmentations, with and without our proposed VFM-based augmentation.

Adding our VFM augmentation on top of the traditional pipeline significantly improves performance across nearly all datasets and models, with the exception of SVOX-Snow, where the baseline already performs strongly. This already shows the potential of using targeted, synthetic training data to address specific visual challenges such as night-time or seasonal changes.

Replacing traditional augmentations with Kornia-based transformations also results in consistent performance gains, confirming the advantage of using more advanced and controlled pixel-level augmentations. Most notably, combining Kornia with our VFM augmentations leads to the strongest results overall, especially on SVOX-Night and Nordland, demonstrating that these two approaches are complementary.

In both cases where our VFM augmentations were applied, whether on top of traditional or Kornia augmentations, performance consistently improved across nearly all evaluation datasets. This highlights the practical value of our VFM-based approach as an effective and scalable method for enhancing ResNet-based VPR methods without requiring changes to the model architecture, particularly when combined with more advanced augmentation pipelines.

B. *Qualitative results*

Figure 4 shows examples where the augmented models correctly found a match in the reference database for SVOX-Night and Nordland, while the original models failed. These cases illustrate how training with our VFM-extened dataset can improve robustness to night-time and winter conditions, particularly in visually challenging scenes with obstacles or low-light conditions.

C. *Ablations*

1) Effect of prompt variation in InstructPix2Pix: To effectively introduce night-time and snow conditions into the training images, an ablation study was conducted to determine which text prompts produced the most visually realistic and consistent results using the InstructPix2Pix model. A wide

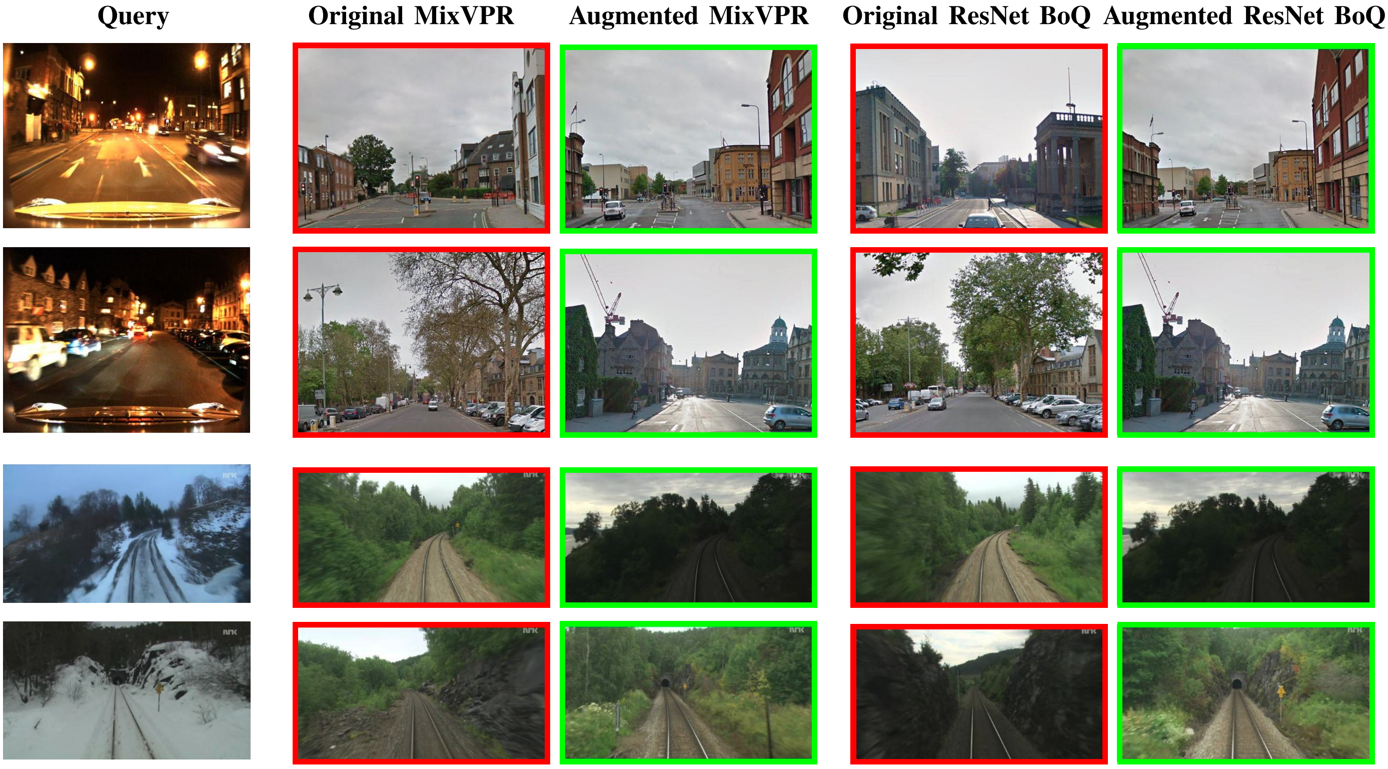

Fig. 4: Qualitative results of predictions of original model compared to the predictions of the augmented models for MixVPR [7] and ResNet BoQ [8] on SVOX-night using night augmented models (first two rows) and on Nordland using winter augmented models (bottom two rows).

range of prompt variations was tested, a selection of which is illustrated in Figure 5.

For the night-time transformation, prompts such as "It is now midnight" and "The sky is now dark and it is night" were evaluated. For the snow transformation, examples included "It is now snowing" and "It is now winter". Each prompt was applied to a diverse set of daytime images, and the resulting outputs were visually inspected to assess realism, semantic consistency, and transformation quality.

Based on qualitative evaluation, the prompt "It is now midnight" consistently yielded the most convincing night-time transformations, while "It is now snowing" produced the most realistic snow-covered scenes.

2) Effect of image and text weights in InstructPix2Pix: After selecting the appropriate prompts for introducing night-time and winter conditions, a follow-up study was conducted to determine the most effective settings for the text guidance and image guidance weights in the InstructPix2Pix model. The goal was to strike a balance between creating the transformation described by the prompt and preserving the semantic and spatial structure of the original image. If the text guidance is too strong or the image guidance too weak, the result can become visually unrealistic or semantically misaligned with the original location. However, overly conservative settings can fail to introduce the desired visual challenge.

To explore this balance, multiple images from a VPR dataset were selected and augmented across a range of values, specifically, text weights from 7.5 to 15.0 and image weights from 1.0 to 1.5. The resulting outputs were manually inspected and qualitatively evaluated based on three criteria: the realism of the transformation, the presence of the intended challenge (night or snow), and the preservation of critical scene elements. Examples of generated outputs using different image weights are shown in Figure 6.

The results of this visual inspection showed that varying the image weight had a greater impact on visual quality than changing the text weight. The default text guidance value of 15.0 provided sufficient adherence to the prompt without over-modifying the image. For night-time augmentation, an image weight of 1.5 yielded the most consistent and realistic results. For winter augmentation, slightly weaker preservation of the original scene was necessary, with an optimal image weight of 1.2. These settings were therefore selected for generating the final synthetic training subsets.

3) Effect of number of diffusion steps in InstructPix2Pix: To identify the most effective number of diffusion steps for generating high-quality augmented images, we conducted an ablation study varying this parameter across a wide range. The number of diffusion steps controls how thoroughly the transformation process refines the image, affecting both visual quality and computational cost. A subset of representative step counts was selected for detailed comparison, as illustrated in Figure 7.

At 5 diffusion steps, image generation was fast, but transformations were often incomplete or visually unconvincing. Night-time prompts led to unrealistic darkening, and snow effects were also rarely realistic, indicating that 5 steps are insufficient for effective scene modification. Using 10 steps

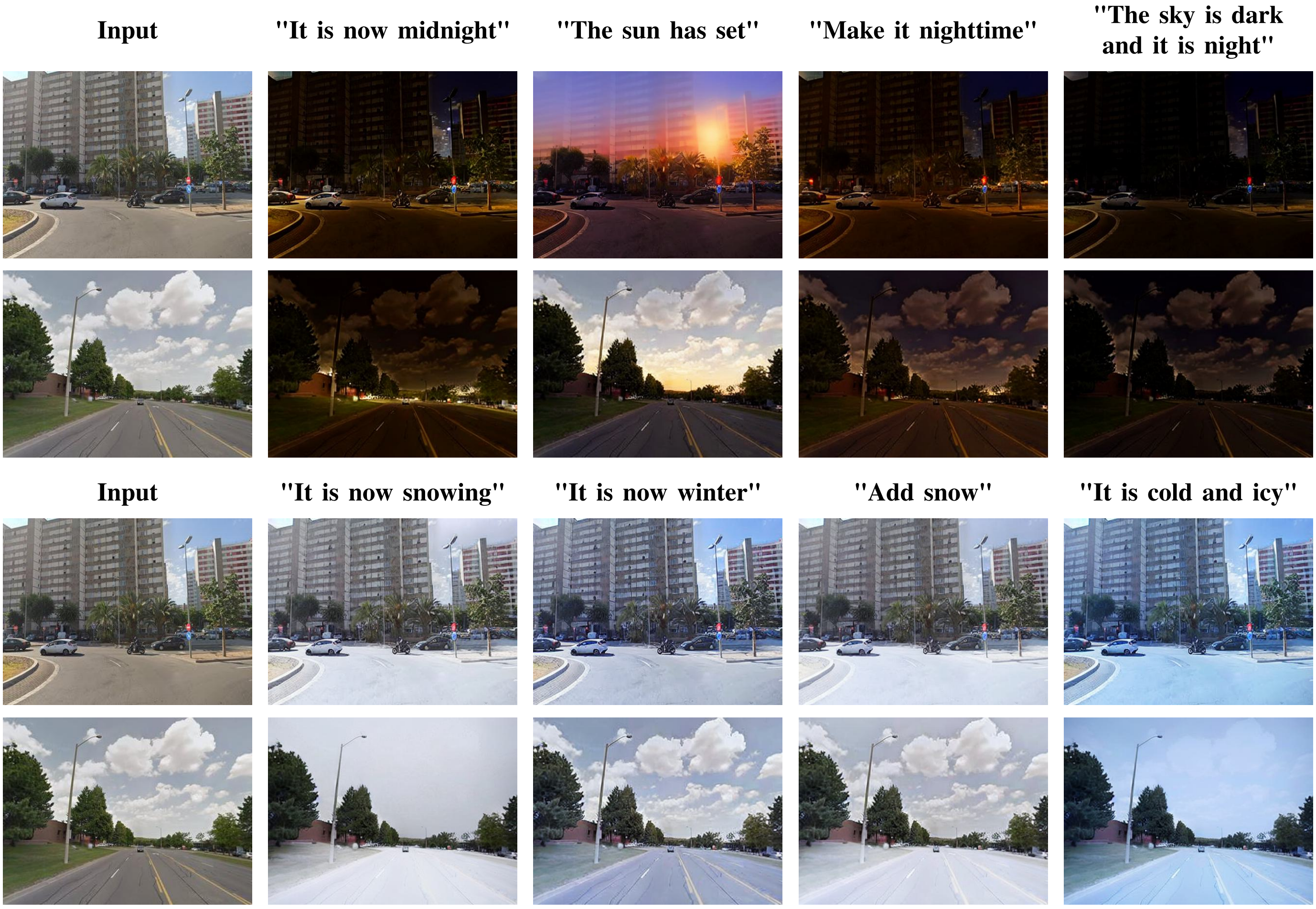

Fig. 5: Generated night (first and second row) or winter (third and fourth row) images with different prompts using InstructPix2Pix [47].

improved quality, with more recognizable night and snow features and better semantic consistency. However, some outputs remained inconsistent, especially in complex scenes. At 20 steps, the model achieved consistently high-quality results. Transformations were realistic and semantically coherent, with well-balanced lighting for night scenes and convincing snow effects. This setting offered the best trade-off between quality and inference time. Increasing to 50 steps gave small visual gains but significantly longer generation times, making it inefficient for large-scale augmentation.

Based on these observations, 20 diffusion steps were identified as the optimal configuration, offering a balance between transformation quality and computational efficiency for training data generation.

4) Effect of augmentation ratio: To introduce new visual challenges using an image-to-image VFM without using too much computation time to generate these new images, this work initially used a training set consisting of the full original GSV-Cities dataset combined with a synthetically augmented subset amounting to one-sixth of the original data added on top of the original dataset. This meant that one in seven images were augmented at training time. This augmentation ratio was chosen as a compromise between dataset extension percentage and generation time. However, it was unclear whether this ratio is optimal in terms of model performance.

To determine the optimal proportion of VFM-augmented images to add to the original training dataset, an ablation study was conducted. The augmentation ratio was varied from 0% to 100% of the original dataset in 16.67% increments (equivalent to one-sixth of the dataset per step). For each configuration, performance was evaluated using MixVPR and ResNet-based BoQ with the results being shown in Table IV.

Augmented Images Added (%)	MixVPR				ResNet BoQ			
	SVOX-Night		Tokyo 24/7		SVOX-Night		Tokyo 24/7	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
0 (original)	63.2	80.7	81.6	91.4	85.5	93.2	**89.9**	94.0
16.67	74.4	86.5	85.1	**93.7**	**89.4**	**95.4**	89.2	94.6
33.33	74.2	85.9	**86.3**	**93.7**	88.9	93.8	88.6	95.2
50	74.2	**87.0**	85.4	93.3	88.7	94.8	89.2	94.9
66.67	75.2	**87.0**	**86.3**	92.7	88.5	**95.4**	**89.9**	**95.9**
83.33	74.9	86.4	85.1	93.3	88.2	94.0	88.3	95.2
100	**75.9**	85.5	85.4	91.7	88.2	93.8	88.9	94.6

TABLE IV: Performance comparison of MixVPR and BoQ ResNet on the SVOX-Night and Tokyo 24/7 datasets for different percentages of augmented night images added to the GSV-Cities training set.

Both models exhibited clear performance gains when 16.67% of the dataset was augmented. However, increasing the proportion of augmented images beyond this point did not

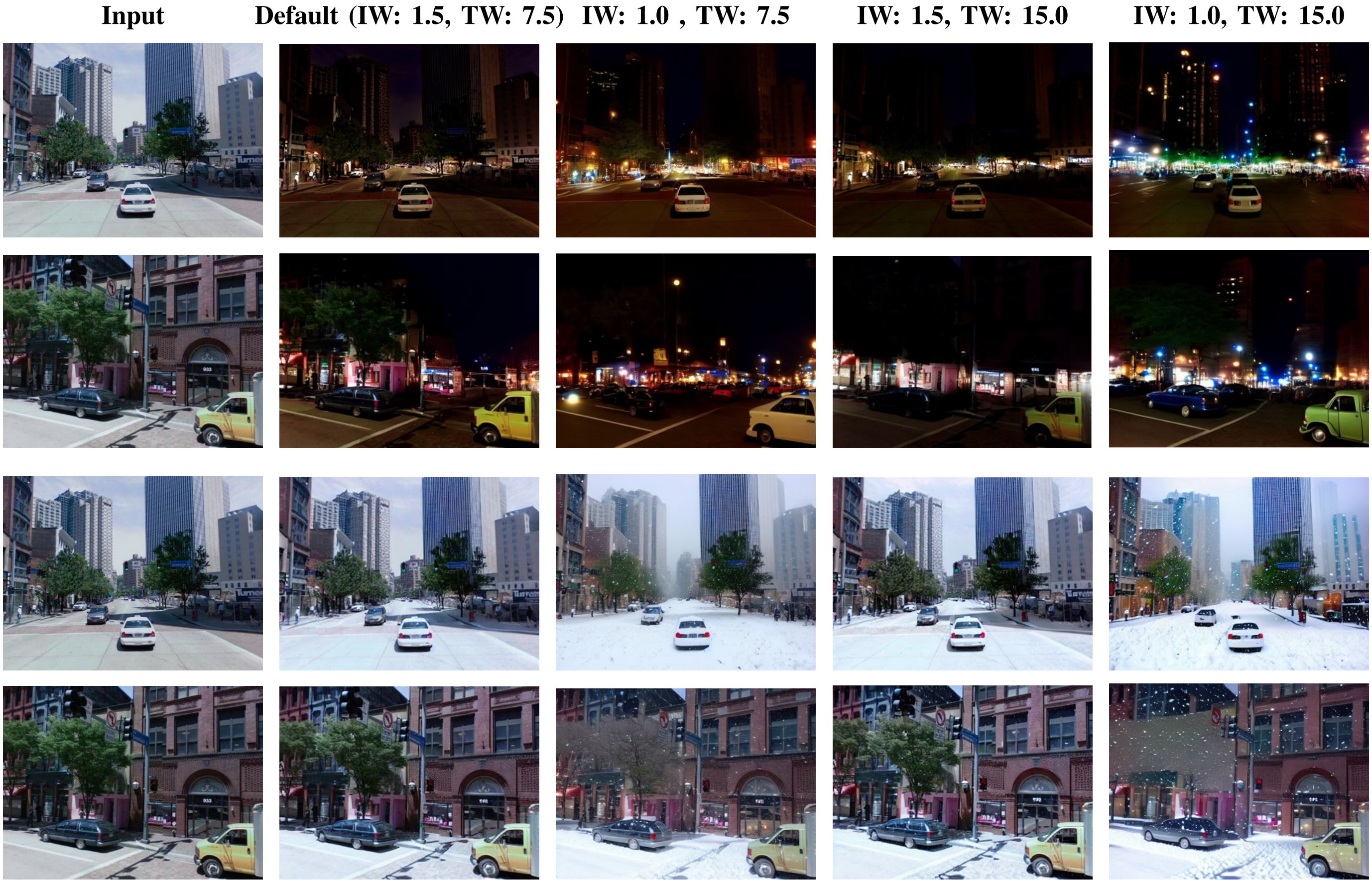

Fig. 6: Generated night (first and second row) or winter (third and fourth row) images with different image weights (IW) and text weights (TW) using InstructPix2Pix [47].

cause significant improvements and, in some cases, resulted in slight performance decrease. This performance plateau suggests limited benefit from further increasing the augmentation ratio.

Given the minimal performance difference beyond the first augmentation step and the high computational cost of generating VFM-based image-to-image augmentations, adding only one-sixth (16.67%) of augmented images to the original dataset offers the most efficient and practical strategy. This ratio provides a strong balance between performance gains and resource usage, and is therefore selected as the optimal augmentation ratio for subsequent experiments.

Dataset	**Model**	**Original model**		**VFM augmented model**			
		R@1	**R@5**	**R@1**	**Δ**	**R@5**	**Δ**
Tokyo 24/7	CricaVPR	94.0	97.5	**94.6**	0.6	**97.8**	0.3
	BoQ DINOv2	**96.5**	**98.1**	**96.5**	0.0	**98.1**	0.0
SVOX-Night	CricaVPR	88.5	95.9	**88.8**	0.3	**96.0**	0.1
	BoQ DINOv2	96.5	**99.5**	**97.7**	1.2	99.4	-0.1
Nordland	CricaVPR	91.0	96.5	**91.6**	0.6	**96.7**	0.2
	BoQ DINOv2	**90.6**	**96.1**	**90.6**	0.0	**96.1**	0.0

TABLE V: Performance of original and VFM night augmented DINOv2-based models on SVOX-Night and Tokyo 24/7 and VFM winter augmented DINOv2-based models on Nordland.

5) Evaluation on DINOv2-based models: While our proposed VFM augmentation strategy has shown clear benefits for ResNet-based models, these architectures no longer represent the true state of the art in VPR. Transformer-based models, particularly those using DINOv2 backbones, have recently become dominant due to their stronger generalization capabilities and performance across diverse conditions. To evaluate whether such models can also benefit from synthetic data augmentation, experiments were conducted using our augmentation pipeline with both CricaVPR and BoQ DINOv2, each trained on the full GSV-Cities dataset with an additional 1/6th night- or winter-augmented subset.

The results are shown in Table V . These results however, indicate that transformer-based models benefit only marginally, if at all, from this form of data augmentation. For BoQ DINOv2, performance marginally increased on SVOX-Night for R@1, but also slightly decreased for R@5. On Tokyo 24/7 and Nordland the results remained unchanged. CricaVPR also showed only minor gains, with R@1 and R@5 increasing slightly on Tokyo 24/7, SVOX-Night and Nordland, but these are changes that are not significant enough to indicate a consistent benefit from training on augmented data.

These findings suggest that DINOv2-based transformer models are less sensitive to the benefits of our VFM dataset extension method. This is likely due to their extensive pretraining on large and diverse image distributions, which may already expose the model to many of the visual challenges simulated by our augmentations.

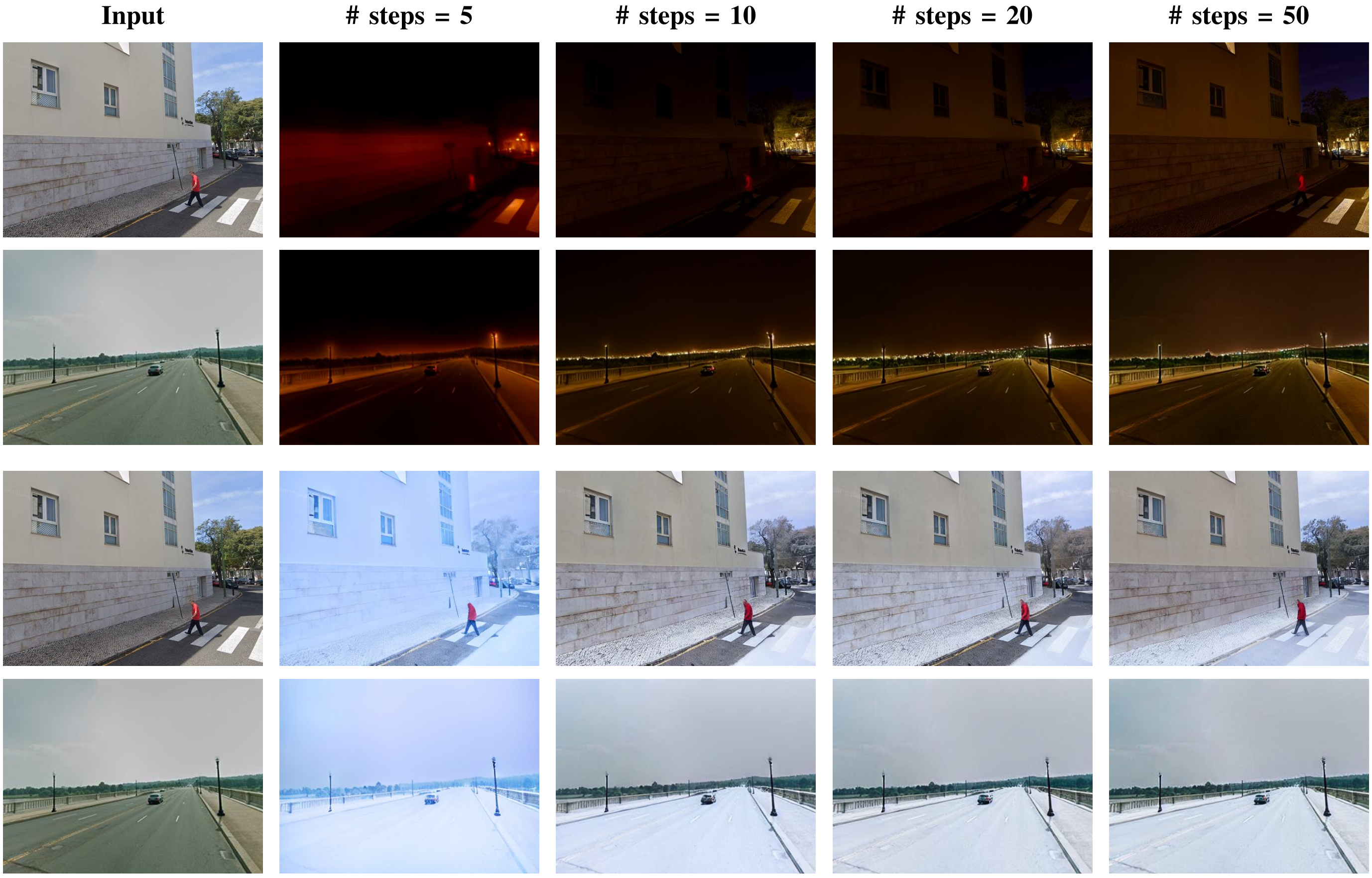

Fig. 7: Generated night (first and second row) or winter (third and fourth row) images with different number of diffusion steps using InstructPix2Pix [47].

VI. Conclusion and future work

This paper introduced a new method for synthetically extending datasets for Visual Place Recognition using Vision Foundation Models. Our goal was to answer the question of how existing datasets can be expanded to improve performance of more efficient ResNet-based VPR models, which are better suited for real-world robotic applications with limited computational resources. Motivated by the limitations of current VPR datasets, we proposed introducing missing visual challenges, like day-to-night and seasonal changes, by generating new image variants from the original data using an image-to-image vision foundation model.

Our results show that, for ResNet-based models, the proposed VFM-based augmentation approach gives significant performance improvements. Replacing traditional augmentations with more advanced Kornia augmentations further enhances results. In both cases where our VFM augmentations were applied, whether on top of traditional or Kornia augmentations, performance consistently improved across nearly all evaluation datasets, with the Kornia+VFM combination achieving the strongest overall results. Our method helps narrow the performance gap between lightweight ResNet architectures and more advanced, but computationally expensive, transformer-based models. We also evaluated the impact of this method on transformer models, which showed only marginal gains or none at all, likely due to their robust backbones already pre-trained on large, diverse datasets that capture many of the same visual variations.

While our method has already shown promising potential, it also opens up many new directions for future research. One interesting research direction would be to explore how these synthetic augmentations can be used more effectively to maximize performance. Additionally, future work could investigate the use of different vision foundation models for data generation, especially as newer and more powerful models continue to emerge. Another interesting research topic would be to explore alternative generative approaches, such as image-to-video or text-to-image models, for producing training data. Lastly, introducing other types of challenges, such as variations in viewpoint, which were not addressed in this study, could possibly further enhance the robustness of VPR models.

In summary, this thesis demonstrates that leveraging the generative power of VFMs to augment training data is a viable strategy for improving VPR systems. By closing the gap between lightweight and high-capacity models, our approach advances the practicality of deploying robust VPR in real-world, resource-constrained environments.

VII. Acknowledgements

I want to thank my supervisors Mubariz Zaffar and Julian Kooij for their help on this project.

References

[1] Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., & Milford, M. J. (2015). Visual place recognition: A survey. *ieee transactions on robotics*, *32*(1), 1–19.

[2] Masone, C., & Caputo, B. (2021). A survey on deep visual place recognition. *IEEE Access*, *9*, 19516–19547.

[3] Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous localization and mapping: Part i. *IEEE robotics & automation magazine*, *13*(2), 99–110.

[4] Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., & Pajdla, T. (2015). 24/7 place recognition by view synthesis. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1808–1817.

[5] Berton, G. M., Paolicelli, V., Masone, C., & Caputo, B. (2021). Adaptive-attentive geolocalization from few queries: A hybrid approach. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2918–2927.

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

[7] Ali-Bey, A., Chaib-Draa, B., & Giguere, P. (2023). Mixvpr: Feature mixing for visual place recognition. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2998–3007.

[8] Ali-Bey, A., Chaib-draa, B., & Giguère, P. (2024). Boq: A place is worth a bag of learnable queries. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17794–17803.

[9] Maffra, F., Teixeira, L., Chen, Z., & Chli, M. (2019). Real-time wide-baseline place recognition using depth completion. *IEEE Robotics and Automation Letters*, *4*(2), 1525–1532.

[10] Zaffar, M., Garg, S., Milford, M., Kooij, J., Flynn, D., McDonald-Maier, K., & Ehsan, S. (2021). Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *International Journal of Computer Vision*, *129*(7), 2136–2174.

[11] Ong, E.-J., Husain, S., & Bober, M. (2017). Siamese network of deep fisher-vector descriptors for image retrieval. *arXiv preprint arXiv:1702.00338*.

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

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[14] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al. (2022). A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, *45*(1), 87–110.

[15] Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K. M., Scherer, S., Krishna, M., & Garg, S. (2023). Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*.

[16] Li, J., Xia, X., Li, W., Li, H., Wang, X., Xiao, X., Wang, R., Zheng, M., & Pan, X. (2022). Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. *arXiv preprint arXiv:2207.05501*.

[17] Mauricio, J., Domingues, I., & Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, *13*(9), 5521.

[18] Warburg, F., Hauberg, S., Lopez-Antequera, M., Gargallo, P., Kuang, Y., & Civera, J. (2020). Mapillary street-level sequences: A dataset for lifelong place recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2626–2635.

[19] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

[20] Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). Orb: An efficient alternative to sift or surf. *2011 International conference on computer vision*, 2564–2571.

[21] Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, 404–417.

[22] Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., & Schmid, C. (2011). Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence*, *34*(9), 1704–1716.

[23] Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. *2010 IEEE computer society conference on computer vision and pattern recognition*, 3304–3311.

[24] Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, *155*, 23–36.

[25] Zhang, X., Wang, L., & Su, Y. (2021). Visual place recognition: A survey from deep learning perspective. *Pattern Recognition*, *113*, 107760.

[26] Simonyan, K. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[27] Chen, Z., Lam, O., Jacobson, A., & Milford, M. (2014). Convolutional neural network-based place recognition. *arXiv preprint arXiv:1411.1509*.

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

[29] Berton, G., Trivigno, G., Caputo, B., & Masone, C. (2023). Eigenplaces: Training viewpoint robust models for visual place recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11080–11090.

[30] Barbarani, G., Mostafa, M., Bayramov, H., Trivigno, G., Berton, G., Masone, C., & Caputo, B. (2023). Are local features all you need for cross-domain visual place recognition? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6155–6165.

[31] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

[32] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763.

[33] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

[34] Izquierdo, S., & Civera, J. (2024). Optimal transport aggregation for visual place recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17658–17668.

[35] Lu, F., Lan, X., Zhang, L., Jiang, D., Wang, Y., & Yuan, C. (2024). Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16772–16782.

[36] Xu, M., Yoon, S., Fuentes, A., & Park, D. S. (2023). A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, *137*, 109347.

[37] Jang, S., & Kim, U.-H. (2023). On the study of data augmentation for visual place recognition. *IEEE Robotics and Automation Letters*.

[38] DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

[39] Chen, P., Liu, S., Zhao, H., & Jia, J. (2020). Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*.

[40] Berton, G., Mereu, R., Trivigno, G., Masone, C., Csurka, G., Sattler, T., & Caputo, B. (2022). Deep visual geo-localization benchmark. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5396–5407.

[41] Wozniak, P., & Ozog, D. (2023). Cross-domain indoor visual place recognition for mobile robot via generalization using style augmentation. *Sensors*, *23*(13), 6134.

[42] Musallam, M. A., Gaudillière, V., & Aouada, D. (2024). Self-supervised learning for place representation generalization across appearance changes. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7448–7458.

[43] Riba, E., Mishkin, D., Ponsa, D., Rublee, E., & Bradski, G. (2020). Kornia: An open source differentiable computer vision library for pytorch. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3674–3683.

[44] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[45] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

[46] Trabucco, B., Doherty, K., Gurinas, M., & Salakhutdinov, R. (2023). Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

[47] Brooks, T., Holynski, A., & Efros, A. A. (2023). Instructpix2pix: Learning to follow image editing instructions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

[48] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

[49] Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., & Irani, M. (2023). Imagic: Text-based real image editing with diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

[50] Kodaira, A., Xu, C., Hazama, T., Yoshimoto, T., Ohno, K., Mitsuhori, S., Sugano, S., Cho, H., Liu, Z., & Keutzer, K. (2023). Streamdiffusion: A pipeline-level solution for real-time interactive generation. *arXiv preprint arXiv:2312.12491*.

[51] Ali-bey, A., Chaib-draa, B., & Giguère, P. (2022). Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, *513*, 194–203.

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

[53] Hausler, S., Garg, S., Xu, M., Milford, M., & Fischer, T. (2021). Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14141–14152.

[54] Wang, X., Han, X., Huang, W., Dong, D., & Scott, M. R. (2019). Multi-similarity loss with general pair weighting for deep metric learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5022–5030.

[55] Vaswani, A. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

[56] Torii, A., Sivic, J., Pajdla, T., & Okutomi, M. (2013). Visual place recognition with repetitive structures. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 883–890.