Visual Place Recognition: Building an Evaluation Framework for Model Robustness

VLAD-IONUȚ GOȘA, University of Twente, The Netherlands

Visual Place Recognition (VPR) is the process of identifying and retrieving images captured at the same location as a given query image. The introduction of VPR pipelines in applications such as autonomous driving and mobile robot localization makes it crucial that models perform image retrieval tasks consistently under challenging conditions such as image blur, lossy compression, or even domain changes such as weather and time of day. Our standardized benchmark compares multiple state-of-the-art VPR pipelines on synthetically generated test datasets to mitigate the effects of uncontrollable variables caused by the image capturing process. Results show how vision transformer backbones are consistently more robust to domain changes and image corruptions compared to traditional convolutional neural network backbones.

Additional Key Words and Phrases: Visual Place Recognition, Robustness, Evaluation Framework, Image Corruption, Image Translation

1 INTRODUCTION

Visual Place Recognition (VPR) is used to denominate the process of figuring out the location of a given query image. It is an essential component for the navigation of mobile robots [12], and autonomous driving [8]. Most approaches present in the literature treat VPR as an image retrieval problem. Assuming that a database of related geotagged images exists, the algorithm is tasked to retrieve the closest matches to the given input. Current state-of-the-art place recognition strategies handle this retrieval process with a two-stage pipeline [20, 35], or as a single-stage pipeline by skipping post-retrieval re-ranking algorithms to reduce latency [14, 5, 2, 19, 3, 16].

(1) Image features are extracted into a compact feature vector. Pairwise comparisons are made between the query vector and every other image in the database. Using either Euclidean distance or cosine similarity, the best matches are returned based on a distance metric.
(2) A post-processing stage refines the results using several re-ranking algorithms.

A critical aspect of VPR systems lies in the ability to identify task-relevant regions in an image. Query images are rarely perfect, indicating that scenes frequently contain distracting elements or significant domain changes such as night and day shifts, temporary occlusions, or even image capture corruptions [4]. The ability of a model to withstand such changes is defined as robustness. In this study, we define the alterations impacting these queries as corruptions. These corruptions can be classified into two main categories: short-term and long-term. Short-term corruptions are represented by common disruptions present in images captured with a mobile camera. We will take a subset of the corruptions proposed by Hendrycks and Dietterich: defocus blur, motion blur, zoom blur, elastic transformations and JPEG compression artifacts [11]. Long-term corruptions are defined as temporary domain shifts caused by weather, season or time of day shifts. Using CycleGAN-Turbo [25], one of the latest works in the field of GAN models, we generated two long-term corrupted datasets: night time and rainy weather starting from a single ground truth subset of the MSLS validation dataset. [32]

TScIT 41, July 5, 2024, Enschede, The Netherlands

2 PREVIOUS WORK

2.1 Robustness in Visual Place Recognition

The need for robust models in VPR is inferred from the applications in which this technology is applied. Previous studies examined the robustness of several state-of-the-art VPR pipelines for daytime and nighttime changes. GaN architectures were used to synthetically generate day-to-night images from an unpaired dataset, removing the need for paired data [22]. The results indicate that using synthetically enhanced datasets yields better results than a less diverse real dataset. However, the paper does not address other potential domain changes, such as weather or seasonal variations, or the potential for generating artificial datasets for robustness evaluation. As our contribution, we show how one-step conditional models such as CycleGAN-Turbo [25] can be used to generate synthetic domain changes that extend beyond day-to-night changes, such as clear to rainy, while also preserving image details without using edge detectors. Fine-tuning was used to improve the robustness of existing architectures against out-of-distribution low-light pictures with promising results [18]. U-Net was used as a normalization layer before a VGG backbone to enhance the feature extraction process, resulting in better descriptors and overall performance [15]. Even though both papers show how different normalization and fine-tuning techniques can increase retrieval performance, they do not show how their optimizations affect a model's performance under other domain changes such as weather. Several other benchmarks have examined CNN and Transformer-based VPR processes on day-night, occlusions, and seasonal changes, albeit, without using synthetic images [31, 2, 6]. We propose an evaluation framework which allows large-scale augmentation of existing datasets, thus enabling researchers to examine a model's robustness under multiple types of short-term and long-term corruptions. To our knowledge, the only previous research that addresses multiple short-term corruptions in VPR applications is Smit's research which is used as a starting point for this study [30].

2.2 Vision Transformers (ViT) and Convolutional Neural Networks (CNN)

Traditionally, CNNs were used as a backbone in VPR due to their ability to extract highly informative feature maps from images. Almost all state-of-the-art architectures use a ResNet-50, ResNet-101 [10] or VGG-16 backbone [29], depending on the available memory

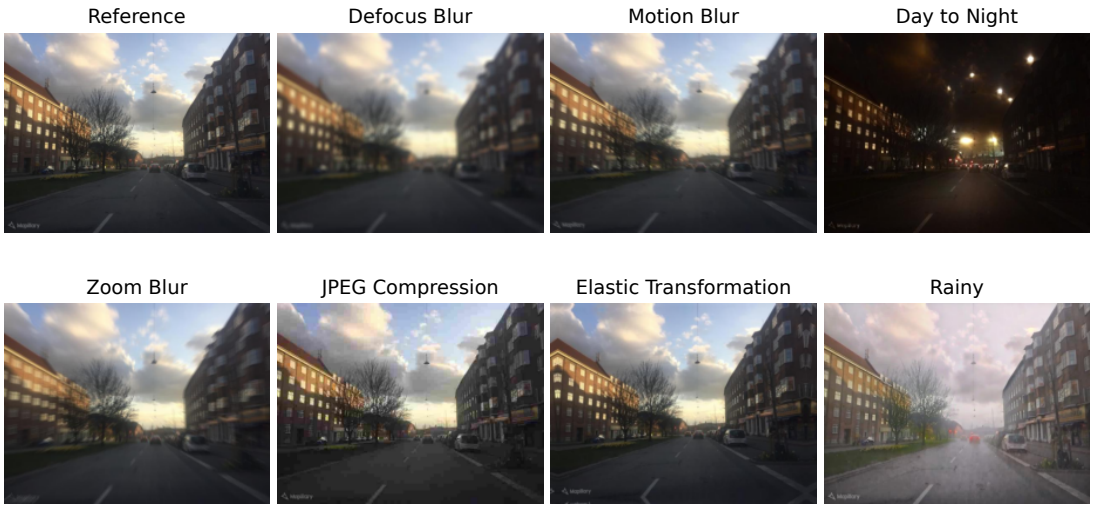

Figure 1. **Corruptions.** Comparison between the original image and the seven proposed short-term and long-term corruptions. All the short-term corruptions presented in the figure are of severity level five. The nighttime and rainy images are generated using CycleGAN-Turbo. The original image is retrieved from the Mapillary Street Level Sequences validation dataset.

and processing resources for training. Most image retrieval pipelines extract local descriptors using CNN backbones, and then use a pooling layer to create a latent space embedding. When a query image is given to such system, it undergoes the same process, and then, based on a distance metric, its distance to the nearest possible matches is computed to retrieve a ranking of the most similar images inside the database. Later revisions of such systems use popular aggregation methods such as GeM [26] and NetVLAD [3] which remained state-of-the-art for VPR applications until the recent advent of vision transformer based architectures [2, 35, 31]. Vision transformers operate by converting images into a series of flattened 2D patches, which are subsequently fed into an encoder along with their respective patch and position embeddings [9]. ViT backbones are utilized in VPR to extract robust local descriptors, which are subsequently combined into a final, global descriptor [13]. A recent study shows the potential of transformer-based architectures to extract powerful descriptors from a small dataset, with recall rates equivalent to or sometimes even better than current CNN-based state-of-the-art approaches. The main proposition is that VPR can be considered a regression task rather than a classification task, with the added benefit of avoiding computationally expensive tasks such as dataset pair-mining and re-ranking [17].

3 PROBLEM STATEMENT

The choice of a representative dataset is an ongoing challenge in the domain of VPR. For training, the most used datasets are Mapillary-SLS [32] and most recently, GSV-Cities [1] due to their extensive coverage of both urban and rural images from all around the globe. The usage of comprehensive datasets ensures that VPR models can derive high-quality descriptors that perform well on novel datasets. However, these datasets do not cover an equally distributed amount of challenging conditions such as weather, season and time of day changes due to the difficulty of capturing such data at a large scale. Image-to-Image translation can bridge this gap by allowing researchers to both augment existing datasets and create corrupted versions of existing validation and test sets. This research will focus on the latter, where we use algorithmically generated corruptions to compute a subset of short-term corrupted images [11], and a CycleGAN-Turbo [25] model to create a subset of long-term corrupted images, mainly clear-to-rainy and day-to-night corruptions. Multiple benchmarks have been proposed to analyze the performance of VPR pipelines. Although most tests share the same MSLS validation dataset as a de facto standard, not all models and image retrieval strategies are tested in the same way among benchmarks. Since image retrieval is a multistage process, even small differences in one component of the process can cause noticeable changes in performance. It is important, therefore, to standardize the process of benchmarking VPR pipelines to ensure consistent points of reference for further research. We combine the processes of data augmentation and model evaluation by extending the standardized benchmarking tool proposed in [6] (See Appendix 6). The proposed framework can be used to augment existing datasets with short-term and long-term corrupted images to both, train or test place recognition pipelines. Compared to previous methods [23, 22], our framework offers a systematic way of augmenting datasets and evaluating VPR pipelines at a larger scale with more diverse corruption types. By augmenting a subset of the MSLS validation dataset, we aim to answer the following research question:

How do state-of-the-art place recognition architectures perform when exposed to corrupted query images?

- What types of corruptions are relevant for place recognition?
- How do transformer-based approaches compare to CNN backbones?
- How do state-of-the-art VPR pipelines perform under queries affected by short-term and long-term corruptions?

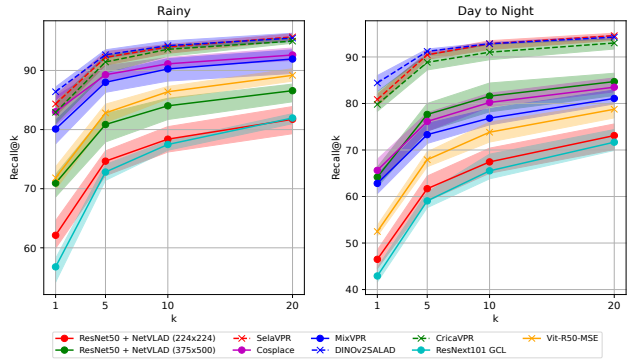

Figure 2. **Margins of error.** Each dot on a line represents the mean recall value for kϵ[1,5,10,20] of a given model starting from left to right. The colored bands show the span between the minimum and maximum value recorded when testing on the five random dataset shuffles.

4 METHODOLOGY

4.1 Corruption Types

Short-Term Corruptions. ImageNet-C [11] represents a natural starting point for generating realistic short-term corruptions. Its collection of corruptions represents a standard benchmark for image classification tasks. Out of 15 types of algorithmically generated corruptions proposed in the paper, only five are used in the benchmark. The first category, blurs, frequently occurs in robotics and autonomous driving because of the swift motion of the cameras mounted on robots and vehicles. *Defocus blur* occurs when an image is out of focus. *Motion blur* is introduced when a camera is moving quickly at the moment of capture. *Zoom blur* occurs when a camera rapidly moves towards an object to be captured. The second category is image compression, represented by *JPEG compression*, which is a lossy compression algorithm often used to reduce file size for low-latency media streaming. The last category is transformations, where *elastic transformation* is used to stretch or contract small regions of an image. Examples of the aforementioned corruptions are shown in Figure 1. Each short-term corruption has five levels of severity, resulting in 25 distinct short-term corruptions (see Appendix 7).

Long-Term Corruptions. Taking inspiration from the Multi-Weather City dataset [23], which used similar techniques to generate various weather, season and temporal domain changes, we use a pre-trained CycleGAN-Turbo model to generate two types of corruptions. *Day to night* domain translations to test a model's robustness against temporal changes and *clear to rainy* to test robustness against weather changes (See Figure 1). CycleGAN-Turbo offers a novel approach to image translation by using a one-step pre-trained text-to-image model capable of generating realistic images. By combining this model with a condition encoder that uses skip-connections, images can be translated to different domains with less artifacts compared to older GAN models. The datasets used to train the generative model are BDD100k [34], which contains diverse driving images under day and night conditions, and DENSE [7] which includes foggy driving images taken from the 'dense-fog' dataset split.

4.2 Evaluation Data

For our evaluation and corrupted dataset generation, we use the Mapillary Street Level Sequences (MSLS) dataset. It is a large-scale heterogeneous dataset containing images from both urban and suburban areas captured in 30 major cities across six continents [32]. Each short-term corruption has 5 corrupted sets, one for each severity level. The long-term corruptions only have one set each. In total, the number of datasets used in one experiment, including the original queries, is 28. To reduce computing overhead, all the corrupted datasets are created from a random sample of 1000 images taken from the MSLS validation split. To ensure that our sampling does not bias the results relative to using the entire validation dataset, we generated five random dataset shuffles of the same size and computed the variance in recall rates across these shuffles for our long-term corruptions, since these tests affect each VPR pipeline's performance the most. The results show the shuffles can indeed affect the recall rates of a given model by 1-3%, but not enough to significantly affect the outcome of our testing (See Figure 2).

4.3 Implementation Details

To ensure the consistency and reproducibility of results, we extend the visual place recognition benchmark proposed by Berton et al. [6] (See Appendix 6), which allows fine-grained control over all the components of a VPR pipeline. A query image is first normalized using the benchmark repository's default mean and standard deviation values. Afterwards, the query image is pre-processed using hard-resizing according to each backbone's supported input dimensions. In the case of transformer-based VPR pipelines, the image is resized to a 1:1 aspect ratio with the height and width divisible by the encoder patch height and width dimensions. Each VPR pipeline's backbone is initialized using their publicly provided weights. VPR approaches that use re-ranking will have an additional step to perform post-processing prediction refinement. The inference batch size is defaulted to 16 images. L2 normalization is applied by default before pooling.

5 EXPERIMENTS AND RESULTS

We conduct a comparative analysis of the baseline ResNet-50 and NetVLAD architecture against state-of-the-art VPR pipelines by assessing their robustness to significant domain changes using long-term corruptions and their consistency in performance against increasing severity levels for each short-term corruption.

5.1 Experimental Setup

Models. For testing, we chose VPR systems that utilize both convolutional neural network (CNN) backbones (NetVLAD [3], CosPlace [5], MixVPR [2] and GCL [16]), and vision transformers (ViT) (SelaVPR [20], DINOv2SALAD [13], CricaVPR [19], and ViT-R50-MSE [17]) to understand better how each type of backbone handles both short-term and long-term corruptions. This list includes the majority of top-performing models on the MSLS validation dataset. All models are pre-trained on either MSLS [32], SF-XL [5] or GSV-Cities [1]. NetVLAD and MixVPR use a ResNet-50 [10] backbone. GCL is tested using a ResNeXt101 [33] backbone. CosPlace uses a ResNet101 backbone. DINOv2SALAD, CricaVPR and SelaVPR all use a pre-trained

Table 1. **Tested models overview.** Each model is pre-trained and tested using their publicly available weights. Wherever possible, instead of using the full 480x640 resolution of the dataset, we use 375x500 (61% of the original resolution) which is a good compromise for geo-localization tasks [6]. All models are tested using their best-performing descriptor dimensions. No PCA was used during testing. * ViT-R50-MSE uses ResNet-50 to extract a feature map which is then projected and flattened to the embedding dimensions required by the ViT model's encoder.

Network	Backbone	Aggregation	Reranking	Features Dim	Training Dataset	Resolution (HxW)	Size (MB)
NetVLAD	ResNet-50	NetVLAD	no	65536	MSLS	224, 224	33.2
NetVLAD	ResNet-50	NetVLAD	no	65536	MSLS	375, 500	33.2
SelaVPR	DINOv2	GeM + Local Adapt	yes	1024	MSLS	224, 224	1363.5
ViT-R50-MSE	ViT-R50	- *	no	768	MSLS	384, 384	374.53
GCL	ResNeXt101-x32d	GeM	no	2048	MSLS	480, 640	331.7
DINOv2SALAD	DINOv2	SALAD	no	8448	GSV-cities	322, 322	335.7
CricaVPR	DINOv2	GeM + Crica Encoder	no	10752	GSV-cities	224, 224	407.3
MixVPR	ResNet-50	MixVPR	no	4096	GSV-cities	320, 320	41.6
CosPlace	ResNet-101	CosPlace	no	2048	SF-XL	375, 500	178.5

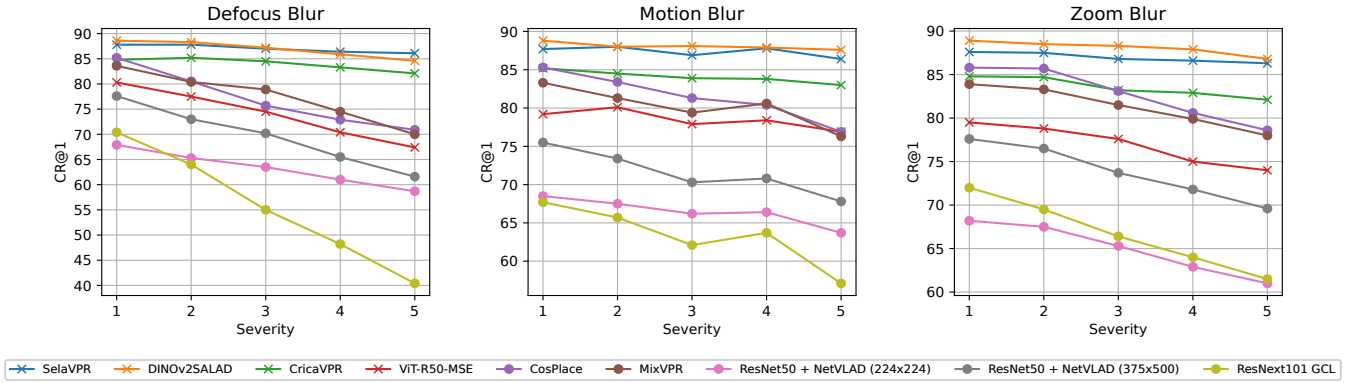

Figure 3. R@1 results obtained by applying blur corruptions to the evaluation dataset. On the y-axis, the corrupted recall (CR@1) is shown, and on the x-axis, the severity level of the corruption is shown; the higher the severity, the heavier the corruption effect is applied to the image. All transformer-based VPR systems are marked with the letter (*x*), whereas the ones that use CNN backbones are marked with a circle (○).

DINOv2 vision transformer backbone. ViT-R50-MSE uses a hybrid transformer with knowledge distilled from a ResNet-50 backbone. An overview of each tested VPR system and its components can be found in table 1.

Metrics. When validating, a database image is deemed a match for a given query if its GPS location falls within a 25-meter radius of the query and the discrepancy in viewing angles is less than 40 degrees. The primary evaluation metric utilized is recall@k (R@k), which represents the proportion of queries in which at least one correct image is selected from its k-nearest neighbours. VPR architectures typically execute similarity searches between global descriptors employing the L2 norm, followed by the Euclidean distance.

5.2 Results

Throughout this section, we investigate how each type of corruption affects the absolute recall values for each model. Specifically, we will first examine the effects of blurring corruptions, focusing on the difference between convolutional neural networks and vision transformers. Secondly, we analyze the overall robustness against compression artifacts and transformations. We then move to the analysis of long-term corruption robustness across all recall rates and the sensitivity of models against considerably large domain changes.

5.2.1 ***Blurring Corruptions.*** Place recognition systems must consider the environmental conditions in which they operate. Blurring is common when dealing with query images captured by moving subjects like robots and autonomous vehicles. The level of blurring depends on the speed of the subject and the camera's characteristics, particularly the shutter speed. To address this, we conduct tests at five different levels of blur severity.

Discussion. Plots of the results are shown in figure 3. Compared to other types of short-term corruptions, such as elastic transformation and compression, blurring effects have a steeper impact on the performance of the models. This effect is explainable by analyzing the drop in performance caused by increasing the severity level. Blurring effects hide essential details in pictures, such as building facades, road markings and foliage. A general trend amongst all the blurring corruption results is the increased robustness of transformer-based networks compared to CNN backbones. While newer aggregation methods such as CosPlace and MixVPR tend to offer competitive performance with DINOv2-based architectures at severity level 1, as the severity level increases, the performance of CNN backbones

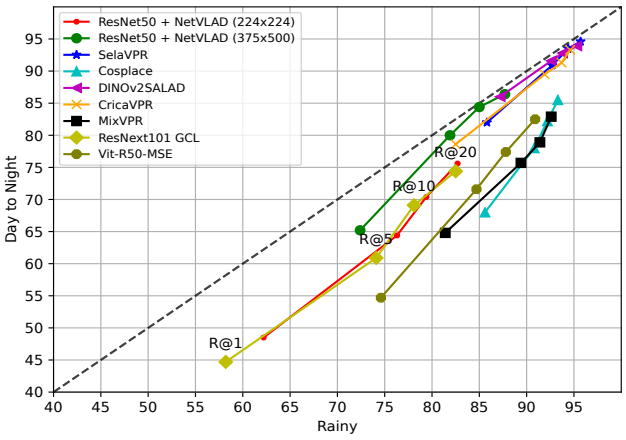

Figure 4. R@k results for k ϵ [1,5,10,20]. The y-axis contains the CR@k values for the night-corrupted images; the x-axis contains the values for the clear-to-rainy corrupted images. Each model is represented by the four recall values, with the leftmost point being CR@1 and the rightmost point on the x-axis being CR@20.

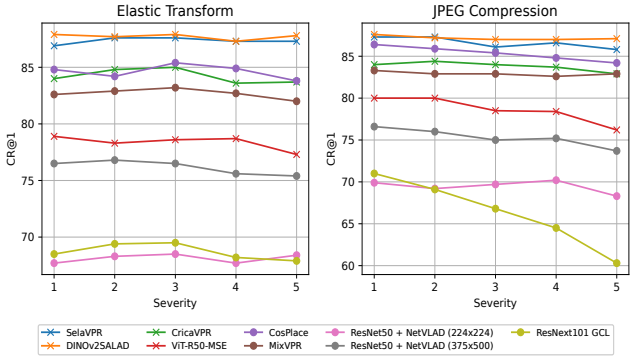

Figure 5. R@1 results obtained by applying elastic transformation and JPEG compression to the evaluation dataset. The formatting of the chart is identical to Figure 3.

generally decreases at an increased rate compared to the DINOv2 VPR pipelines. In general, MixVPR and CosPlace offer the best performance compared to other CNN backbones; however, it should be noted that MixVPR is based on a ResNet-50 backbone, which is a considerably lighter variant compared to CosPlace's ResNet-101. For the vision transformers category, DINOv2SALAD is generally the top-performing architecture, closely followed by SelaVPR, which uses re-ranking. Vit-R50-MSE's performance is more similar to the CNN top performers MixVPR and Cosplace than to its DINOv2 transformer counterparts. The difference can be attributed to the much more extensive pre-training process used in DINOv2 compared to the one used in Vit-R50. ResNeXt101-GCL is competitive with the computationally more expensive NetVLAD aggregation running on a lower 224x224 resolution under motion and zoom blurs, indicating that GCL could be a viable alternative when defocus blurs are not common. Moreover, these results demonstrate that various training approaches, such as regression in the case of Vit-R50-MSE, as well as classification in CosPlace, can generate robust descriptors without the need for expensive pair-mining commonly used in contrastive learning techniques.

5.2.2 ***Compression and Transform Corruptions.*** Various VPR applications necessitate specialized cameras to manage the conditions the subjects face. For instance, vehicles and robots might need gimbal systems to stabilize their camera on all six degrees of freedom due to the inertia caused by fast movement changes. One consequence caused by small shifts in the perspective of the camera is elastic transformation. JPEG compression, conversely, arises from a lossy compression technique employed to decrease the image file size before sending it over a network.

Discussion. The results of this experiment are presented in figure 5. Elastic transformations and JPEG compressions do not significantly degrade a model's performance as the severity increases. This indicates that CNN's and transformer backbones are generally quite robust against these types of corruptions. The only exception that can be noted is ResNeXt101 GCL, whose performance follows a downward trend under JPEG compression as the severity level increases. It should be mentioned that GCL achieves its best performance using PCA, but we did not utilize it in our experiments. MixVPR and CosPlace show compelling results, being on par with their transformer-based counterparts in JPEG compression and closely behind SelaVPR and DINOv2SALAD in elastic transformation. Vit-R50-MSE performs better than ResNet-50 at 375x500, but achieves lower results than MixVPR.

5.2.3 ***Long-term Corruptions.*** An important aspect of outdoor VPR applications is the ability of a model to recognize places under more difficult domain changes. While datasets such as MSLS offer sub-tasks such as summer to winter to test such challenges, we find that the image pairs often contain inconsistencies, including slightly different fields of view that could affect the experiment's results. Our test includes a novel way of handling this test by generating images using a GaN model, which ensures that the field of view of our images remains constant after the domain translation. We test two domain changes, day to night and clear to rainy weather, which are some of the most common corruption cases in outdoor VPR images.

Discussion. The results of our experiment can be seen in figure 4. The presented chart places each long-term corruption on one axis and a dashed line along y=x. A model performs equally well on both domain changes if the recall rates follow the reference y=x line. If the performance of one model is higher on one domain change than the other, then its results will be closer to that corruption's axis. The results indicate that all models perform better against the rainy corruption. This effect is explained by the fact that the night-time generated images tend to shadow out important details in the pictures, whereas the rainy corruption is less severe in this regard. DINOv2-based VPR pipelines consistently outperform their CNN-based counterparts on both types of corruption. Moreover, DINOv2-based pipelines perform consistently well across both domain changes because their results are closer to the standard y=x line. DINOv2SALAD is the best-performing model at R@1, closely

Table 2. **R@1 results.** The "Clean" column contains the non-corrupted validation set results. The average R@1 value across all severity levels is presented for short-term corruptions. For long-term corruptions, as there are no severity levels, the actual R@1 values are displayed. Each horizontal line divides the models based on their training dataset. The first section is trained on MSLS, the second on GSV-Cities, and the third on SF-XL. All models marked with a * symbol are transformer-based. The best-performing ViT-based VPR system results are bolded, whereas the best CNN-based system results are italicized.

Model	Clean	Defocus Blur	Motion Blur	Zoom Blur	Elastic Transform	JPEG Compression	Rainy	Night
NetVLAD (224x224)	67.4	63.3	66.5	65.0	68.1	69.5	62.2	48.5
ResNeXt101 GCL	72.4	55.6	63.3	66.7	68.7	66.3	58.2	44.7
NetVLAD 375x500	76.8	69.6	71.6	73.8	76.2	75.3	72.4	65.2
ViT-R50-MSE*	80.3	74.0	78.5	77.0	78.4	78.6	74.6	54.7
SelaVPR*	87.1	87.0	87.4	87.0	87.3	86.6	85.8	82.0
CricaVPR*	84.9	84.0	84.1	83.5	84.2	83.8	82.5	78.6
MixVPR	84.0	77.5	80.2	81.3	82.7	82.9	81.4	64.8
DINOv2SALAD*	**88.1**	**87.2**	**88.1**	**88.1**	**87.7**	**87.2**	**87.2**	**86.0**
CosPlace	*86.6*	*78.7*	*81.5*	*82.8*	*84.6*	*85.3*	*85.6*	*68.0*

followed by SelaVPR with re-ranking. The results for R@5, R@10 and R@20 are very similar among all transformer-based networks. MixVPR and CosPlace achieve the highest results in the rainy corruption with R@1 values of 81.4 and 85.6, respectively. However, they achieved lower results in the night-generated images, with R@1 values of 64.8 for MixVPR and 68 for CosPlace. NetVLAD achieves good results using the higher resolution 375x500 images, with R@1 values of 72.4 for the rainy corruption and 65.2 for the night corruption. It is interesting to analyze how NetVLAD clusters can generalize well on night-time images compared to the other, newer CNN-based approaches. ResNext101 with GCL offers comparative results with ResNet50+NetVLAD on R@10 and R@20; however, it cannot achieve the same performance on R@1 and R@5 without applying PCA. Vit-R50-MSE is slightly ahead of ResNet50+NetVlad 375x500 on the rainy corruption, but it achieves lower overall results on the night corruption.

5.2.4 ***R@1 Robustness Overview.*** The experiments performed during this research captured recall rates for values of kϵ[1,5,10,20]. Given the extensive amount of data collected, we will focus solely on analyzing the R@1 values to demonstrate each model's capability to accurately retrieve an image captured at the same location from the database. The results can be found in table 2.

Discussion. The best R@1 performance is achieved by the DINOv2SALAD architecture. It achieves impressive results under all scenarios, including the night-time long-term corruption where the loss compared to the clean dataset is only 2.4%. It is closely followed by SelaVPR, which uses re-ranking, an expensive post-processing stage. CricaVPR achieves solid results at a lower resolution compared to DINOv2 and a bigger descriptor size. ViT-R50-MSE achieves lower results than CricaVPR, but it surpasses the performance of ResNet-50 at 375x500 in all corruptions except the long-term nighttime corruption. However, these results show how regression-based hybrid architectures can surpass the performance of models with considerably larger descriptor dimensions trained on the same dataset. For CNN-based architectures, CosPlace is the best performing method. It achieves comparative results with ViT-based pipelines on multiple tests, including the rainy long-term corruption. However, its performance on the night corruption is 13.5% lower than CricaVPR, which achieved the second lowest results out of the four tested ViT pipelines, further highlighting the advantage of the pre-trained DINOv2 backbone in these systems. Compared to other CNN backbones, MixVPR achieves high results using a ResNet-50 backbone, showing the advantages the MixVPR all-MLP aggregation method offers.

A common pattern highlighted by these results is the consistent robustness of DINOv2-based backbones against corruptions that hide essential details of the picture. A potential explanation for these results could be attributed to the large-scale pre-training of the DINOv2 foundational model compared to the other backbones present in the tested VPR systems. A foundation DINOv2 [24] model is pre-trained on 142 M images compared to the 1.2 M images used in IMAGENET-1K [27], the model used to pre-train the ResNet-50 foundation model.

6 CONCLUSION AND FUTURE WORK

This study expands upon previous works to introduce a novel robustness testing methodology for VPR pipelines that generates reproducible, fast and comparable results. Expanding upon the work of Berton's deep visual benchmarking tool [6] and Smit's work on short-term corruption evaluation for VPR [30], we create an automated way of generating short-term and long-term corruptions for any available VPR dataset. Moreover, the flexibility offered by the benchmarking framework enables researchers to perform ablation studies to further review the performance of their models against commonly occuring image corruptions. We believe that our experiments showcase a new perspective upon VPR pipeline testing, one that is crucial for creating robust models that are able to perform well on out-of-distribution data. Firstly, we show how DINOv2-based networks are consistently more robust against all types of corruptions compared to their CNN counterparts. Secondly, we discover an interesting pattern in the results of all blur corruptions where DINOv2-backed pipelines' performance decreases much slower compared to CNN's. Thirdly, we observe how newer approaches to VPR such as MixVPR and CosPlace offer much more compelling results using newer, more efficient aggregation methods compared to older implementations such as NetVLAD. Lastly, we show how other training approaches such as regression in the case of ViT-R50-MSE, as well as classification in CosPlace can achieve similar to state-of-the-art performance on multiple corruptions with backbones trained on less data compared to DINOv2. Along with

all the results, we showcase a novel approach to robustness testing in VPR by creating and evaluating the performance of VPR systems under GaN generated long-term corruptions. In this final section, we will discuss about the limitations of our testing, the choices of corruptions and future work that could be applied to further expand upon this study.

6.1 Limitations

As Berton et al. acknowledge in their deep visual geo-localization benchmark paper, despite the benchmark's modularity, the framework is clearly focused on VG methods used in outdoor urban environments. However, we believe that the framework's versatility offers a good ground for expansion, especially for robustness testing where the focus is placed on the corruption-generation aspect. For our experiments, this framework offered a standardized testing ground to ensure reproducible results.

All the experiments were conducted on one Nvidia A16 GPU; therefore, due to the lack of time and the number of tests needed to gather the required data, we only tested using a subset of the queries available in the MSLS validation dataset. Out of 11084 suitable queries, we sampled a random shuffle of 1000 queries which we used to create all the testing datasets. We agree that the choice of a random sample instead of the full validation set can introduce bias, such as consistently favouring some models against others. To account for the statistical fluctuations of the results, we analyzed in Fig. 2 the maximum deviation of the long-term corruption results based on five different shuffles of 1000 images. The chart shows how for each model, the possible error rate is between 1-3%. Moreover, it highlights the effects of random dataset sampling over the absolute retrieval performance of each model.

6.2 Corruption choices

Our framework purposefully omits some of the ImageNet-C corruptions. In this subsection, we will discuss why each corruption type was excluded, and what kinds of alternatives our framework offers. Firstly, gaussian, shot and impulse noises were removed since the chances of one query image being corrupted with a very severe case of such noise are relatively rare. The snow, frost and fog corruptions were not used due to our long-term corruption inclusion. We argue that the effects applied by ImageNet-C do not affect the domain of the image, but it merely applies several filters and perturbations which do not portray realistic weather and seasonal changes. To address this issue, we proposed the usage of long-term corruptions generated using GaN architectures, which allow for more realistic domain changes that portray more plausible query images. Time limitations restricted our ability to train and test a CycleGAN-Turbo model using summer-to-winter domain change to replace the snow and frost corruptions completely. Therefore, we encourage future researchers to approach this challenge in the next section. The last two corruptions, brightness and contrast were not included since they portray a very specific change instead of a combination of variables that can affect one image. However, our framework can be extended to include these types of short-term corruptions if needed. The short-term corruptions included in the framework were chosen because we focused on testing the various blurs commonly present in query images captured by moving subjects, as well as elastic transformations caused by the usage of different camera lenses. JPEG compression was chosen since realistically, some VPR pipelines could run on a separate server, and lossy compression is one of the most common ways to reduce network latency.

6.3 Future Work

Long-term corruptions. Our long-term corruption generation is based on two pre-trained CycleGAN-Turbo models. The results consistently show how transformer-based VPR pipelines consistently outperform their CNN counterparts, while handling the two long-term corruptions, rainy and nighttime, equally well. As future work, we propose an extension on our corruptions set, mainly summer to winter domain translation using datasets such as Oxford RobotCar [21] which includes various weather and seasonal changes, including heavy snowy weather, or Extended CMU seasons which depicts urban, suburban and park scenes from the city of Pittsburgh under varying seasonal conditions [28].

Model testing. Our testing methodology unavoidably affects some models such as ResNeXt101 with GCL due to our choice of not using PCA during testing. Due to time and hardware constraints, we decided to keep our methodology strict, therefore to test each model using only one descriptor size, at the resolution suggested by their authors. Moreover, this study focuses on the overall robustness of each model without taking into consideration the differences in inference times, FLOPs and overall efficiency of each VPR system. We propose a clearer separation based on each model's descriptor size, inference speed and model size in future robustness frameworks in order to avoid unfair comparisons.

Pre-training effects. One argument used in this paper is that DINOv2-based pipelines are inherently more robust to all types of corruptions due to the extensive amount of data used in the pre-training process of the foundation model. We suggest an extensive overview on how the amount of data and diversity in the training dataset used by the foundation model correlates with the robustness of VPR pipelines.

REFERENCES

[1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. "GSV-Cities: Toward Appropriate Supervised Visual Place Recognition". In: *Neurocomputing* 513 (Nov. 2022), pp. 194–203. ISSN: 09252312. DOI: 10.1016/j.neucom.2022.09.127. arXiv: 2210.10239[cs].

[2] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. *MixVPR: Feature Mixing for Visual Place Recognition.* Mar. 3, 2023. DOI: 10.48550/arXiv.2303.02190. arXiv: 2303.02190[cs].

[3] Relja Arandjelović et al. *NetVLAD: CNN architecture for weakly supervised place recognition.* May 2, 2016. DOI: 10.48550/arXiv.1511.07247. arXiv: 1511.07247[cs].

[4] Giovanni Barbarani et al. *Are Local Features All You Need for Cross-Domain Visual Place Recognition?* Apr. 12, 2023. DOI: 10.48550/arXiv.2304.05887. arXiv: 2304.05887[cs].

[5] Gabriele Berton, Carlo Masone, and Barbara Caputo. *Rethinking Visual Geo-localization for Large-Scale Applications.* Apr. 7, 2022. DOI: 10.48550/arXiv.2204.02287. arXiv: 2204.02287[cs].

[6] Gabriele Berton et al. *Deep Visual Geo-localization Benchmark.* June 9, 2023. DOI: 10.48550/arXiv.2204.03444. arXiv: 2204.03444[cs].

[7] Mario Bijelic et al. *Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather.* June 30, 2020. DOI: 10.48550/arXiv.1902.08913. arXiv: 1902.08913[cs].

[8] Dzung Doan et al. "Scalable Place Recognition Under Appearance Change for Autonomous Driving". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV).* 2019 IEEE/CVF International Conference on Computer Vision

(ICCV). Seoul, Korea (South): IEEE, Oct. 2019, pp. 9318–9327. ISBN: 978-1-72814-803-8. DOI: 10.1109/ICCV.2019.00941.
[9] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. June 3, 2021. DOI: 10.48550/arXiv.2010.11929. arXiv: 2010.11929[cs].
[10] Kaiming He et al. *Deep Residual Learning for Image Recognition*. Dec. 10, 2015. DOI: 10.48550/arXiv.1512.03385. arXiv: 1512.03385[cs].
[11] Dan Hendrycks and Thomas Dietterich. *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations*. Mar. 28, 2019. DOI: 10.48550/arXiv.1903.12261. arXiv: 1903.12261[cs,stat].
[12] Michael Horst and Ralf Möller. "Visual Place Recognition for Autonomous Mobile Robots". In: *Robotics* 6.2 (June 2017). Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, p. 9. ISSN: 2218-6581. DOI: 10.3390/robotics6020009.
[13] Sergio Izquierdo. *serizba/salad*. original-date: 2023-11-21T08:57:31Z. May 8, 2024.
[14] Sergio Izquierdo and Javier Civera. *Optimal Transport Aggregation for Visual Place Recognition*. Nov. 27, 2023. DOI: 10.48550/arXiv.2311.15937. arXiv: 2311.15937[cs].
[15] Tomas Jenicek and Ondřej Chum. *No Fear of the Dark: Image Retrieval under Varying Illumination Conditions*. Aug. 23, 2019. DOI: 10.48550/arXiv.1908.08999. arXiv: 1908.08999[cs].
[16] María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. *Generalized Contrastive Optimization of Siamese Networks for Place Recognition*. Apr. 20, 2023. arXiv: 2103.06638[cs].
[17] María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. *Regressing Transformers for Data-efficient Visual Place Recognition*. Jan. 29, 2024. DOI: 10.48550/arXiv.2401.16304. arXiv: 2401.16304[cs].
[18] Bingxi Liu et al. *NocPlace: Nocturnal Visual Place Recognition via Generative and Inherited Knowledge Transfer*. Mar. 21, 2024. arXiv: 2402.17159[cs].
[19] Feng Lu et al. *CricaVPR: Cross-image Correlation-aware Representation Learning for Visual Place Recognition*. Apr. 1, 2024. arXiv: 2402.19231[cs].
[20] Feng Lu et al. *Towards Seamless Adaptation of Pre-trained Models for Visual Place Recognition*. Apr. 3, 2024. DOI: 10.48550/arXiv.2402.14505. arXiv: 2402.14505[cs].
[21] Will Maddern et al. "1 year, 1000 km: The Oxford RobotCar dataset". In: *The International Journal of Robotics Research* 36.1 (Jan. 1, 2017). Publisher: SAGE Publications Ltd STM, pp. 3–15. ISSN: 0278-3649. DOI: 10.1177/0278364916679498.
[22] Albert Mohwald, Tomas Jenicek, and Ondřej Chum. *Dark Side Augmentation: Generating Diverse Night Examples for Metric Learning*. Sept. 28, 2023. DOI: 10.48550/arXiv.2309.16351. arXiv: 2309.16351[cs].
[23] Valentina Musat et al. "Multi-weather city: Adverse weather stacking for autonomous driving". In: *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, BC, Canada: IEEE, Oct. 2021, pp. 2906–2915. ISBN: 978-1-66540-191-3. DOI: 10.1109/ICCVW54120.2021.00325.
[24] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. Feb. 2, 2024. DOI: 10.48550/arXiv.2304.07193. arXiv: 2304.07193[cs].
[25] Gaurav Parmar et al. *One-Step Image Translation with Text-to-Image Models*. Mar. 18, 2024. DOI: 10.48550/arXiv.2403.12036. arXiv: 2403.12036[cs].
[26] Filip Radenović, Giorgos Tolias, and Ondřej Chum. *Fine-tuning CNN Image Retrieval with No Human Annotation*. July 10, 2018. arXiv: 1711.02512[cs].
[27] Olga Russakovsky et al. *ImageNet Large Scale Visual Recognition Challenge*. version: 3. Jan. 29, 2015. DOI: 10.48550/arXiv.1409.0575. arXiv: 1409.0575[cs].
[28] Torsten Sattler et al. "Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, June 2018, pp. 8601–8610. ISBN: 978-1-5386-6420-9. DOI: 10.1109/CVPR.2018.00897.
[29] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Apr. 10, 2015. DOI: 10.48550/arXiv.1409.1556. arXiv: 1409.1556[cs].
[30] Peter Smit. "Visual Place Recognition under Image Corruptions". In: ().
[31] Ruotong Wang et al. *TransVPR: Transformer-based place recognition with multi-level attention aggregation*. Apr. 13, 2022. DOI: 10.48550/arXiv.2201.02001. arXiv: 2201.02001[cs].
[32] Frederik Warburg et al. "Mapillary Street-Level Sequences: A Dataset for Lifelong Place Recognition". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, June 2020, pp. 2623–2632. ISBN: 978-1-72817-168-5. DOI: 10.1109/CVPR42600.2020.00270.
[33] Saining Xie et al. *Aggregated Residual Transformations for Deep Neural Networks*. Apr. 10, 2017. arXiv: 1611.05431[cs].
[34] Fisher Yu et al. *BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning*. Apr. 8, 2020. DOI: 10.48550/arXiv.1805.04687. arXiv: 1805.04687[cs].
[35] Sijie Zhu et al. *R^{2}Former: Unified Retrieval and Reranking Transformer for Place Recognition*. Apr. 6, 2023. DOI: 10.48550/arXiv.2304.03410. arXiv: 2304.03410[cs].

A AI DISCLOSURE

During the preparation of this work the author used Grammarly in order to spell-check the contents of this paper. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the work.

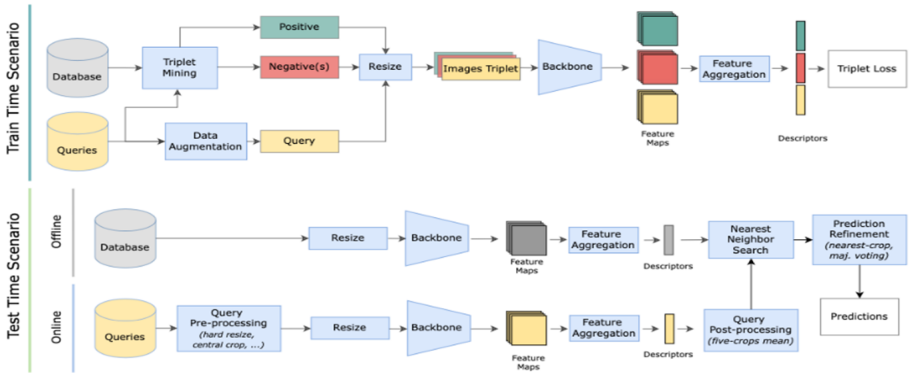

Figure 6. VPR Pipeline benchmark proposed by Berton et al [6]. Each framework component can be interchanged to allow consistent testing and evaluation procedures.

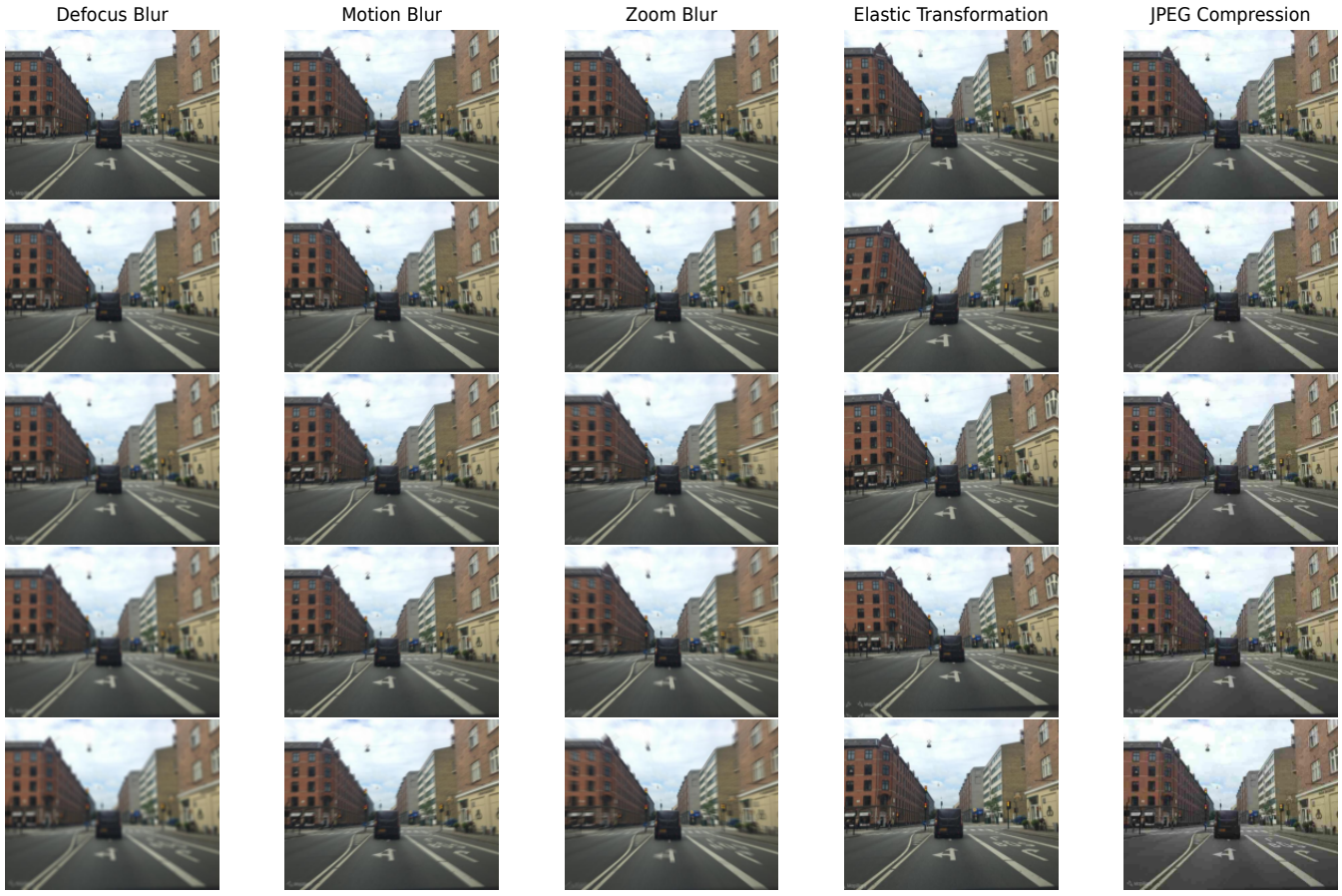

Figure 7. **Short-term corruptions with severities.** Each column contains a short-term corruption category. Severity increases top to bottom, from level 1 to 5.