CRITICAL REVIEW

Night-to-Day Image Translation for Retrieval-based Localisation

Sachith M. Gunawardane

Contents

1 Introduction

The research delves into the challenging task of visual localisation or image comparison across varying illumination conditions, such as weather and lighting changes. This has significant implications, particularly in applications like robotics, where determining a robot's position and orientation in its environment is crucial for navigation and mapping purposes. This involves continuously tracking one's position relative to previous positions and the surrounding environment.

Traditionally, stable approaches to visual localisation have leveraged image retrieval techniques. These techniques involve identifying the most similar image to a query photo from a database of geo-tagged images. However, when dealing with vastly different illumination conditions, such as transitioning from day to night, existing deep neural models struggle due to a lack of suitable training data, specifically labelled image pairs. The challenge lies in the arduous task of collecting hundreds or thousands of images from diverse positions and conditions.

To address this challenge, the research explores the use of image-to-image translation, employing GANs, with unpaired (unlabeled) data. This approach involves transforming the visual properties of images from one domain to resemble those of another domain, where domains are defined solely by collections of data (images). Importantly, this process operates in an unsupervised manner, avoiding the need for extensive labelled datasets.

The specific domain chosen for this research is images captured from car-mounted cameras, with a focus on supporting autonomous driving applications. The authors' model, ToDayGAN, not only alters the visual properties of the target domain but also addresses the properties crucial for image comparison tools. This innovative approach holds promise for improving visual localisation across diverse illumination conditions, advancing the capabilities of autonomous systems in real-world scenarios.

2 Comparison with Related Work

The ToDayGAN model's architecture delves into several key areas, prompting the author to investigate existing work within these domains. These areas include,

- Image-to-Image Translation
- Place Recognition and Localisation
- Image Translation for Visual Location.

2.1 Image-to-Image Translation

Regarding Image-to-Image Translation, traditional GANs generate images by sampling from a probability distribution. However, with translation approaches like CycleGAN, the output is conditioned on a specific input. Zhe et al. [1] introduced CycleGAN as an unsupervised framework for image-to-image translation, aiming to learn to translate an image from one domain (X) to another (Y) without paired examples. This involves learning a mapping G: X → Y such that the distribution of G(X) is indistinguishable from Y, using adversarial loss and cycle consistency loss (ensuring F(G(X)) ≈ X). This method, illustrated in Figure 1, operates on unpaired data, assuming an underlying relationship between domains despite the absence of paired input-output examples.

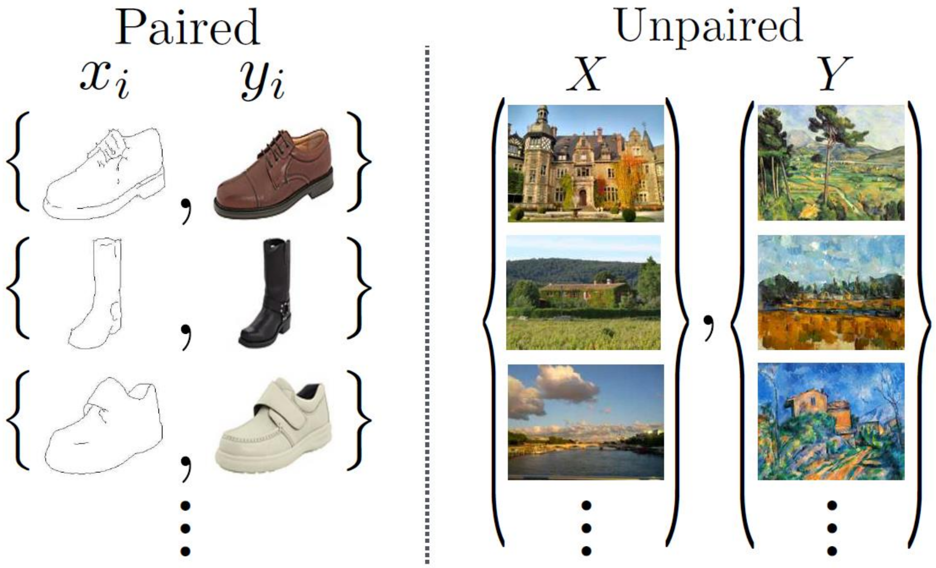


Figure 1: Example of Paired and Unpaired Datasets as [1]

The researchers' assumption of an underlying relationship between domains, even without paired input-output examples, is illustrated in Figure 1 through unpaired samples. These samples are seen as distinct renderings of the same fundamental world, prompting the researchers to aim at learning this relationship. Despite the lack of supervision at the level of paired examples, the researchers exploit supervision at the data domain level. However, using conditional GANs for this purpose does not inherently ensure that individual inputs and outputs (X and Y) are paired meaningfully.

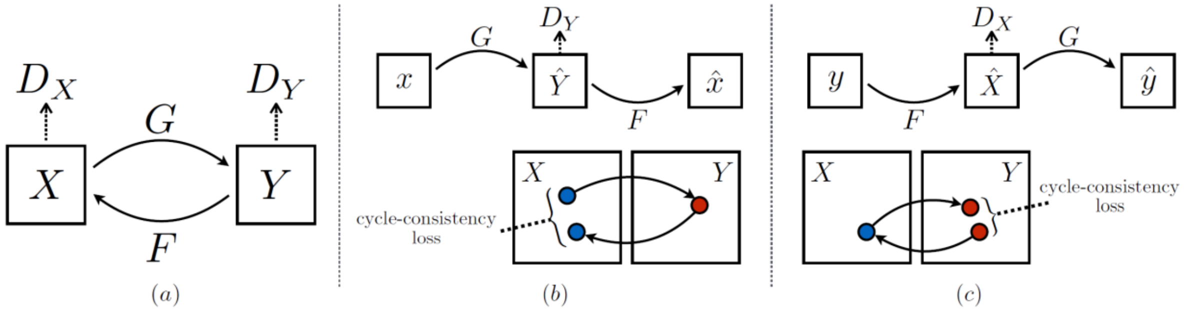


Figure 2: High-level Architecture of CycleGAN

Furthermore, it's noted that optimising the adversarial objective in isolation can be challenging and often leads to the well-known problem of mode collapse, where all input images map to the same output image, stalling optimisation progress. CycleGAN introduces simple forward-backwards consistency to mitigate this issue, a strategy commonly employed in models deployed in the language domain. Figure 2 depicts the CycleGAN's architecture, showcasing a model with two mapping functions and forward/backward cycle-consistency losses for domains X and Y. (a) is the model with two mapping functions, (b) forward cycle-consistency loss considering domain X, (b) backward cycle-consistency loss considering domain Y.

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) = &\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y)\\ &+ \mathcal{L}_{\text{GAN}}(F, D_X, Y, X)\\ &+ \lambda\mathcal{L}_{\text{cyc}}(G, F),\end{aligned}$$

Figure 3: CycleGAN's Full Objective Function

The CycleModel undergoes training using both an adversarial loss and a cycle consistency loss, as illustrated in Figure 3, which outlines the objective function. The cycle consistency loss serves to regularise the highly unconstrained task of unidirectional image translation within the CycleGAN framework. This regularisation helps maintain consistency and coherence in the translated images, preventing erratic or non-intuitive transformations. Figure 4 illustrates the output generated by CycleGAN.

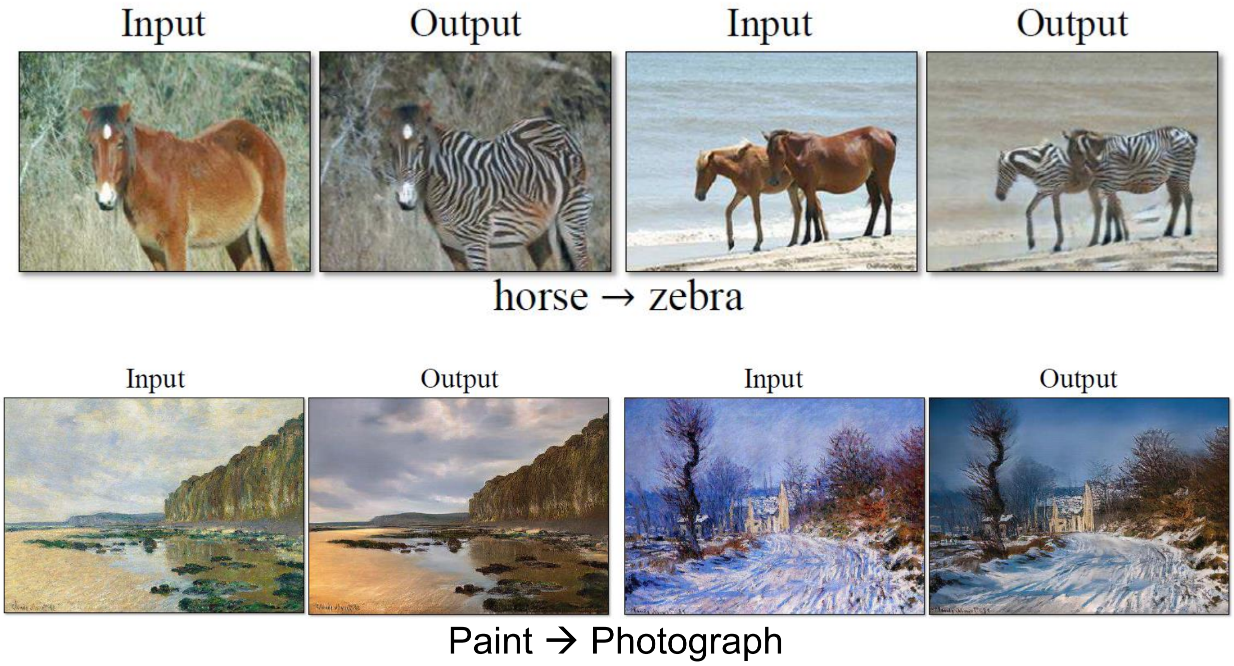


Figure 4: CycleGAN's Output

The applications of CycleGAN are diverse and include tasks such as style transfer, object transfiguration, attribute transfer, and photo enhancement. However, despite its versatility, CycleGAN may encounter challenges and failures due to the distribution characteristics of the training datasets. For instance, Figure 5 highlights a failure in style transfer where the person also mistakenly transformed into a zebra. This failure can be attributed to the model being trained on datasets like ImageNet, which lack images depicting a person riding a house or a zebra, leading to erroneous transformations in specific contexts.

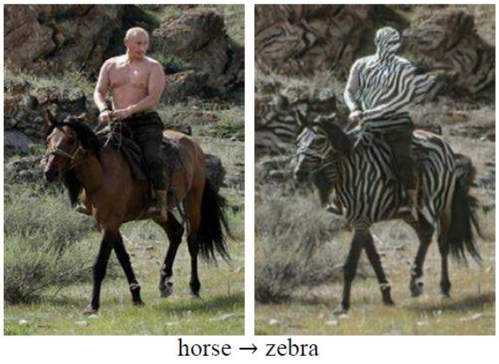


Figure 5: CycleGAN's Failure Scenario

CycleGAN's scalability is limited due to the interconnectedness of its networks, which are jointly tied to two specific domains, typically denoted as A and B. Introducing another domain, say C, necessitates the addition of four new networks: A to C, C to A, B to C, and C to B. To address this limitation, Anoosheh et al. [2] proposed a novel model known as ComboGAN, which decouples the domains and networks from each other. ComboGAN's architecture divides the generator networks used in CycleGAN into halves, labelling the frontal halves as encoders and the latter halves as decoders. This modification expands the model's scope by incorporating multiple generators and discriminators, enabling more nuanced and realistic transformations across different domains. Figure 6 illustrates the architecture of ComboGAN translation of one domain to all other domains.

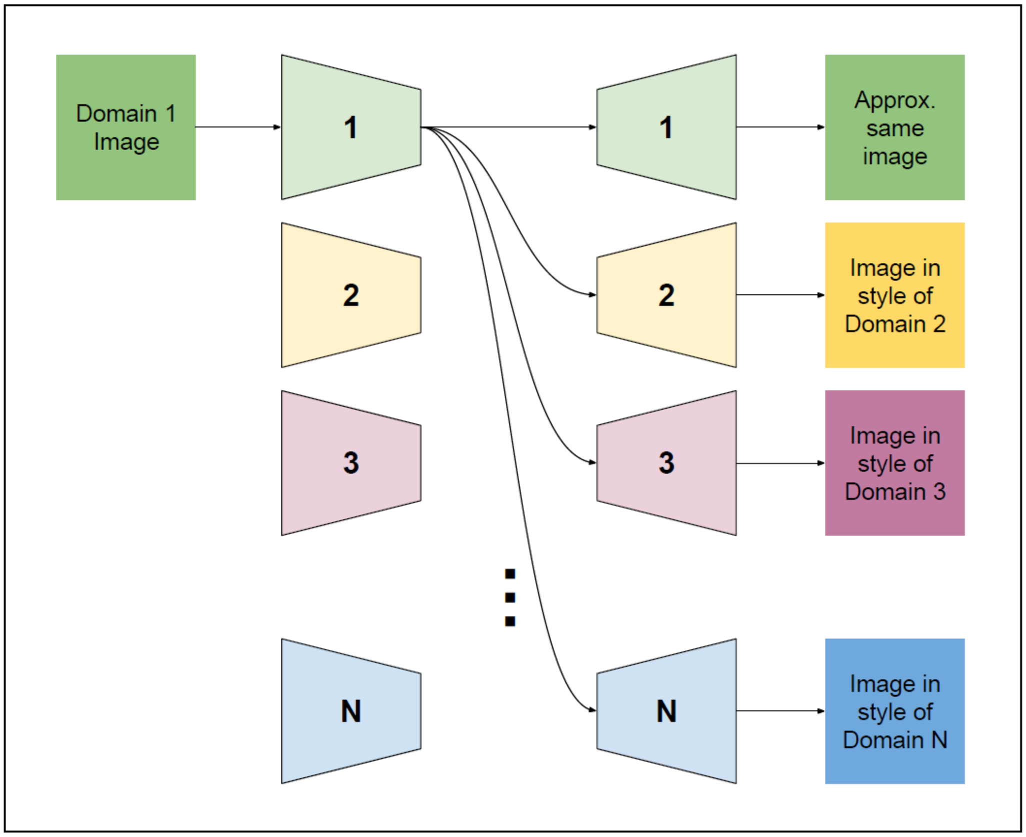

Domain 1 Image
1
2
3
N
1
2
3
N
Approx. same image
Image in style of Domain 2
Image in style of Domain 3
Image in style of Domain N

Figure 6: Example inference functionality of translation from one domain to all others

At its core, ComboGAN comprises multiple generator and discriminator networks, each dedicated to a specific domain. This design allows for simultaneous translation between multiple pairs of domains, facilitating a richer and more comprehensive transformation process. Leveraging these interconnected networks, ComboGAN excels in capturing intricate details and preserving the semantic content of images during translation, as depicted in Figure 7, which aligns the architecture of encoders and decoders with generators and discriminators.

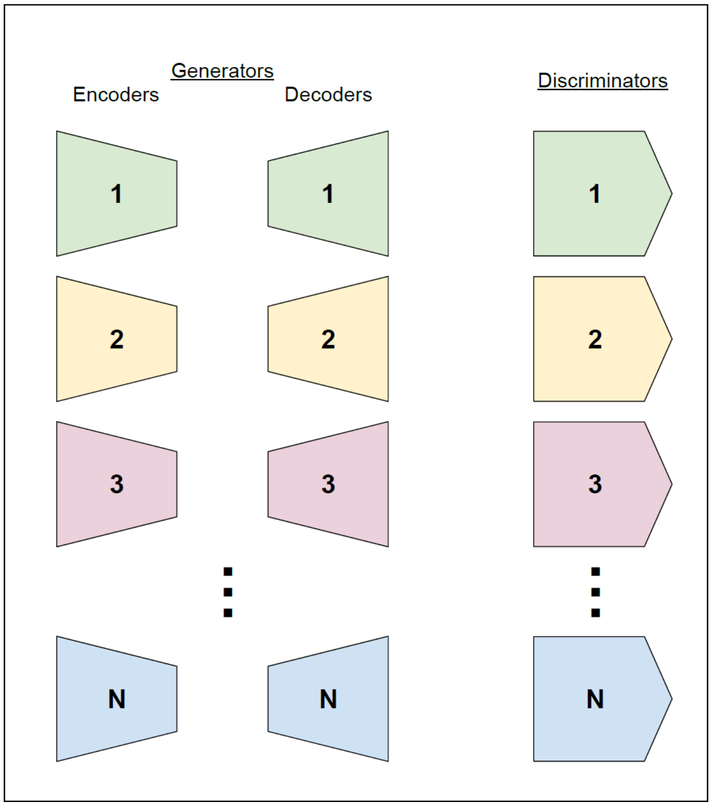


Figure 7: Model design setup for N domains

A notable advantage of ComboGAN is its ability to generate diverse and high-quality outputs. The presence of multiple generators encourages the exploration of varied translation possibilities, resulting in a broader spectrum of results. This diversity enhances the fidelity and creativity of the translated images, making ComboGAN particularly effective in applications requiring versatile and realistic image transformations. Figure 8 showcases output examples from ComboGAN, where original images lie on the diagonal.

In various domains such as art, fashion, and design, ComboGAN finds extensive utility for seamlessly converting images between different styles, textures, and aesthetics. This capability empowers artists, designers, and creators to experiment and innovate with visual content effortlessly. Additionally, ComboGAN's architecture enables fine-grained control over the translation process, allowing users to tailor transformations according to specific preferences and requirements.

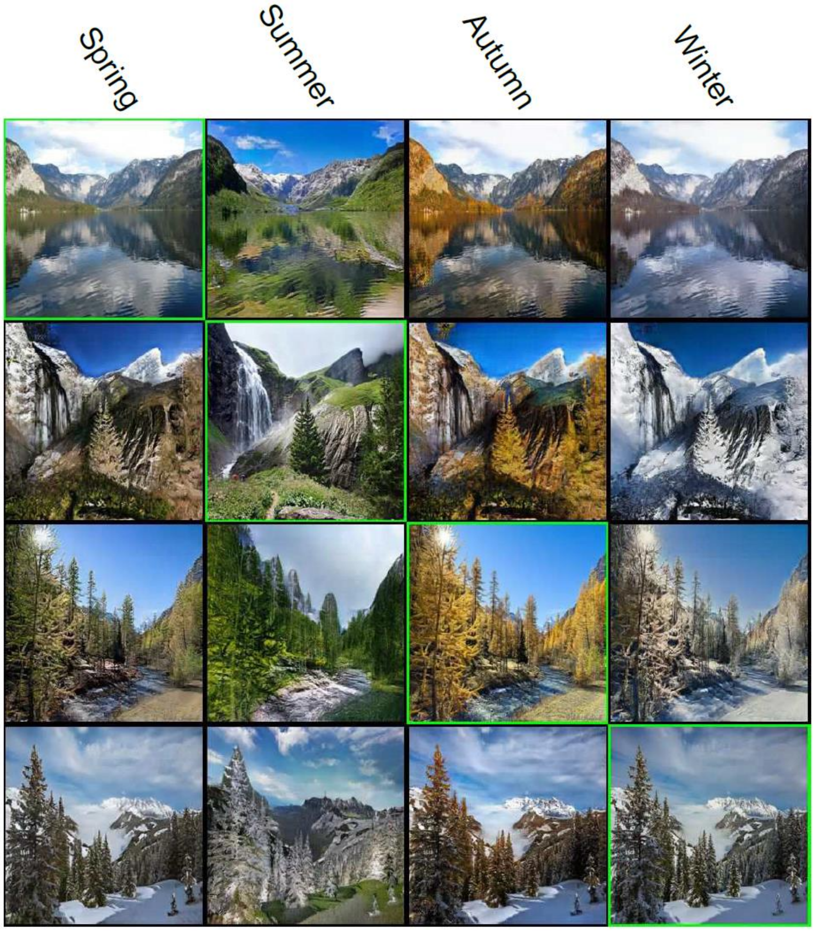


Figure 8: Output of ComboGAN

In cases involving two domains, ComboGAN operates equivalently to CycleGAN, maintaining the underlying foundation without necessitating changes. However, achieving the desired results requires the encoders to place input images into a shared representation when dealing with more than two domains. This latent space ensures that all inputs are equally suitable for any domain transformation, with decoders refilling them with domain-specific details. Achieving this shared representation implies that encoders learn to mask qualities that distinguish images among domains, leaving decoders to reintroduce domain-specific characteristics.

2.2 Place Recognition and Localisation

Place recognition involves identifying real-world locations from images, essentially performing location classification. Ideally, these processes should be invariant to image properties such as camera location, orientation, and weather conditions. In contrast, visual localisation focuses on determining the camera's position relative to a local environment or a global map.

A widely used tool in place recognition is the VLAD descriptor, short for Vector of Locally Aggregated Descriptors. This descriptor plays a pivotal role in the recognition process by extracting local visual features from images representing various places. These features encompass elements like edges, corners, and textures, providing a detailed representation of the visual content. After extracting these features, the VLAD descriptor performs an aggregation step, grouping similar features based on their visual characteristics. This step is crucial as it reduces redundancy and captures the essence of the visual content more efficiently than simply listing all features individually.

The aggregated features are then represented as a vector, encapsulating collective information about the visual features specific to a particular place. Each element within this vector corresponds to a specific aspect or pattern found in the images, aiding in differentiating between places based on visual cues. During place recognition tasks, the VLAD descriptor compares the visual features of a new image with stored VLAD descriptors of known places. This comparison involves measuring the similarity between these vectors, allowing the system to determine whether the new image corresponds to a previously identified place or location.

DenseVLAD and NetVLAD represent advancements in computer vision, extending the capabilities of the traditional VLAD method. DenseVLAD enhances VLAD by introducing dense sampling, which involves extracting features densely across the entire image rather than focusing on sparse key points. This approach results in richer feature vectors that offer a more comprehensive and detailed description of the visual content within an image. DenseVLAD is especially beneficial for tasks that require nuanced and detailed feature representations.

In contrast, NetVLAD takes a different route by integrating VLAD into a neural network framework. It employs a learnable aggregation layer called the NetVLAD layer, which dynamically learns the aggregation process during training. This adaptability enables NetVLAD to optimise feature aggregation based on specific tasks or datasets, enhancing performance and flexibility in handling diverse visual data. Both DenseVLAD and NetVLAD build upon the foundational concepts of VLAD, further refining and augmenting its capabilities in capturing and representing visual information effectively.

DenseVLAD's strength lies in its ability to capture rich and detailed features across the entire image, offering a more comprehensive view of visual content. On the other hand, NetVLAD leverages the power of neural network architectures to achieve adaptable and optimised feature aggregation, making it versatile and efficient in handling complex visual data.

2.3 Image Translation for Visual Localisation

Domain shift and dataset bias are commonly encountered challenges in methods that learn from data. These issues arise when a model trained on a particular dataset is applied to data that differ in specific characteristics, leading to performance degradation. To address these challenges, domain adaptation techniques are employed to mitigate the effects of domain shifts.

Image translation for visual localisation is a crucial aspect of addressing domain shifts in image-based localisation tasks, especially when dealing with different visual conditions. This is particularly evident when images captured under varying illumination conditions significantly impact localisation accuracy.

The utilisation of CycleGAN in image translation plays a pivotal role in overcoming domain shifts caused by differing illumination conditions. CycleGAN enables the visual transformation of images from one domain to another, such as converting night-time images to daytime equivalents before conducting feature matching. This approach effectively bridges the gap between images captured under diverse lighting conditions, enhancing the robustness and accuracy of image-based localisation systems.

3 Methodology

The Night-to-Day Image Translation for Retrieval-based Localization research delves into the challenges of domain shift and dataset bias commonly faced in image-based localisation tasks. In particular, the study highlights the significant impact of varying illumination conditions on localisation accuracy. To address this, the researchers employ CycleGAN, a powerful image translation tool, to visually translate images from one domain to another, such as converting night-time images to their daytime counterparts before feature matching.

Furthermore, the research utilises DenseVLAD as the image featurisation tool, emphasising its superiority over modern methods like NetVLAD in generalisation for day-night image-matching tasks. This choice underscores the importance of robust and effective feature extraction methods in overcoming domain shifts and improving localisation accuracy across diverse visual conditions.

Moreover, the image translation model utilised in the study is based on ComboGAN, a versatile framework that can accommodate multiple domains automatically if necessary. However, the researchers tailor ComboGAN's setup to suit their specific problem, modifying its discriminators for improved performance in the localisation task using translated images.

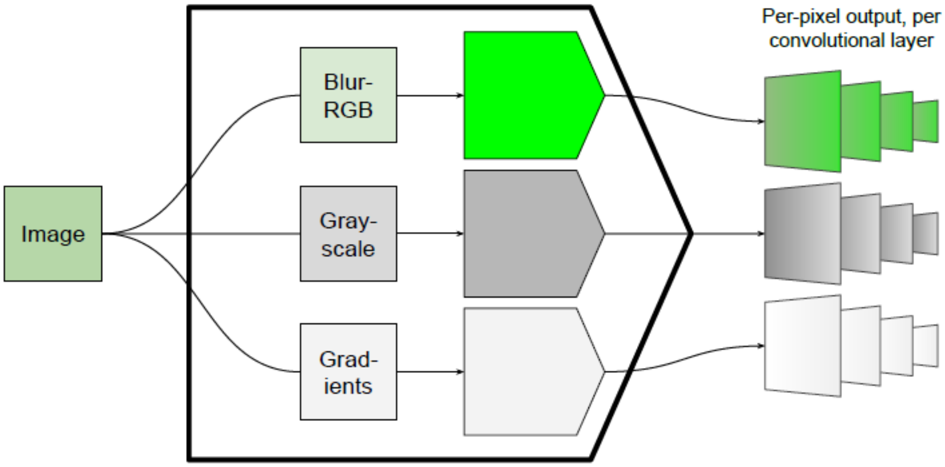


Figure 9: ToDayGAN's discriminator architecture in [3]

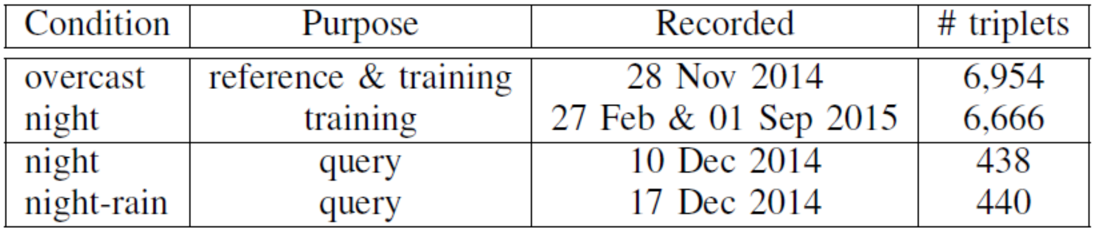
Figure 9 provides insights into the modifications made to the discriminators within the ToDayGAN setup. Each discriminator comprises three network clones operating independently on the input image's blurred RGB, grayscale, and xy-gradients. This modification allows the discriminators to focus separately on texture, colour, and gradients, enhancing their ability to capture diverse visual features. These discriminators, with equal architecture and hyperparameters, collectively contribute to the improved performance observed in the localisation task using translated images.

3.1 Dataset

The research leveraged the extensive Oxford RobotCar dataset [4], which comprises multiple video sequences capturing a 10km route in Oxford, England, from an autonomous vehicle's perspective. The dataset encompasses diverse environmental conditions using three Point Grey

Grasshopper2 cameras mounted on the vehicle's left, right, and rear sides. These traversals spanned a year, offering variations in lighting, time of day, and weather conditions, resulting in a massive collection of over 20 million high-resolution images (1024x1024 pixels).

The research selected a subset of specific traversals from this vast dataset for the experiments. These subsets consisted of image triplets representing corresponding left, right, and rear views at each timestamp. The research utilised the RobotCar Seasons variant to evaluate their approach, which provides accurate camera poses for reference and query images. The reference set, "Day" or "overcast-reference", consisted of 6,954 camera triplets captured during daytime conditions.



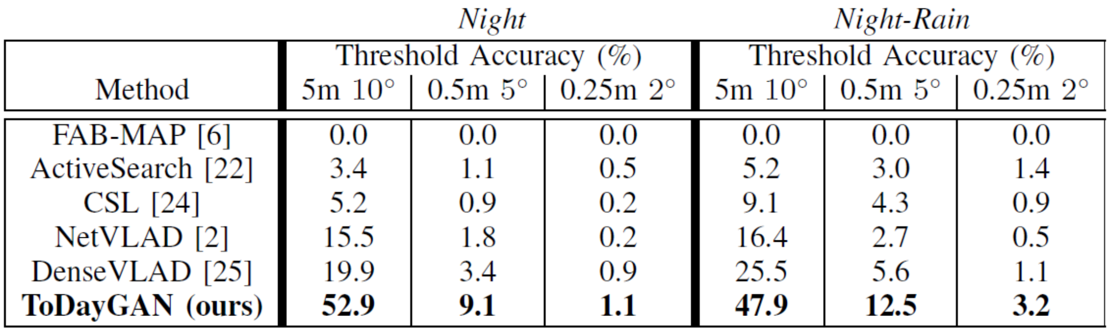
Condition	Purpose	Recorded	# triplets
overcast	reference & training	28 Nov 2014	6,954
night	training	27 Feb & 01 Sep 2015	6,666
night	query	10 Dec 2014	438
night-rain	query	17 Dec 2014	440

Figure 10: RobotCar Seasons dataset

Additionally, the research utilised two query sets: one comprising 438 triplets captured at night ("night") and another with 440 triplets captured at night during rain ("night-rain"). The latter's purpose was to test the transferability of the technique, trained without rain conditions, to a visually distinct domain. Furthermore, a training set of 6,666 night-time triplets ("Night-train") from other traversals of the RobotCar dataset was used exclusively for model training. Figure 10 provides a summarised overview of these subsets in a table format, facilitating easy extraction of information.

4 Results and Analysis

The proposed ToDayGAN method represents a significant leap forward in visual localisation, particularly for night-time query images matched against a daytime reference set with known poses. Figure 11 provides a comprehensive overview of the achieved results, highlighting key metrics such as pose accuracy thresholds. Notably, ToDayGAN demonstrates remarkable performance improvements compared to previous state-of-the-art methods. For instance, at the standard 5m 10° threshold, ToDayGAN achieves an impressive 52.9% accuracy, marking a substantial 2.65x enhancement over the previous best-performing method, DenseVLAD, which scored 19.9% accuracy. Even at the stricter 0.5m 5° threshold, ToDayGAN achieves 9.1% accuracy compared to DenseVLAD's 3.4%, showcasing a proportional boost in performance. Moreover, ToDayGAN showcases robustness beyond day/night transitions, performing well even on the more challenging Night-Rain query set with 47.9% accuracy at the 5m 10° threshold.



	Night			*Night-Rain*		
	Threshold Accuracy (%)			Threshold Accuracy (%)		
Method	5m 10°	0.5m 5°	0.25m 2°	5m 10°	0.5m 5°	0.25m 2°
FAB-MAP [6]	0.0	0.0	0.0	0.0	0.0	0.0
ActiveSearch [22]	3.4	1.1	0.5	5.2	3.0	1.4
CSL [24]	5.2	0.9	0.2	9.1	4.3	0.9
NetVLAD [2]	15.5	1.8	0.2	16.4	2.7	0.5
DenseVLAD [25]	19.9	3.4	0.9	25.5	5.6	1.1
ToDayGAN (ours)	**52.9**	**9.1**	**1.1**	**47.9**	**12.5**	**3.2**

Figure 11: Overall results comparison

An ablation study conducted as part of the research sheds light on the effectiveness of different components within the ToDayGAN architecture. Notably, using separate discriminators for colour, luminance, and gradients yields a substantial improvement over alternative approaches, such as employing a single discriminator or concatenating inputs. Including relativistic discriminator loss is beneficial in stabilising training and enhancing results. Additionally, the dual evaluation procedure involving flipping and re-evaluating each query further refines accuracies, while an increase in image resolution from 286x286 to 512x512 demonstrates noticeable gains.

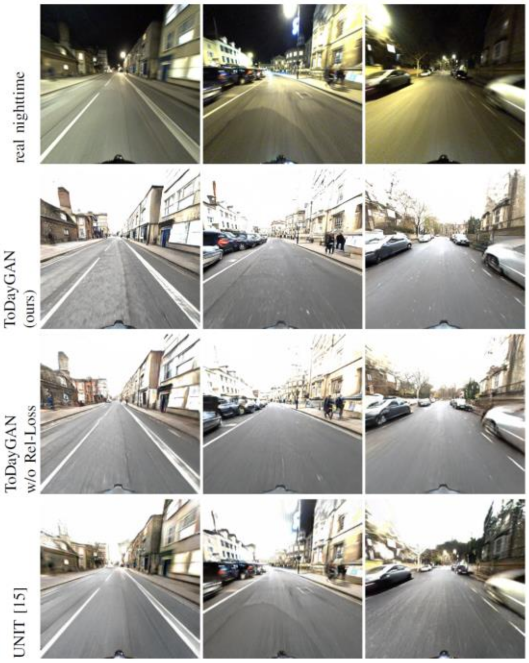


Figure 12: ToDayGAN and UNIT Output

Figure 12 complements these findings by illustrating real night images before translation, synthetic day images generated using the best-performing ToDayGAN model, synthetic day images produced by the same model but without the Relativistic Loss, and synthetic day images generated by the UNIT model. The research also compares ToDayGAN with other generative models like CycleGAN and UNIT, showcasing ToDayGAN's superiority in producing higher-quality and more localisable translated images, particularly for capturing fine details.

The authors delve into insights regarding the success of certain modifications within ToDayGAN, such as employing separate gradient discriminators and multi-scale output. Splitting discriminators allow for specialisation in different low and high-level characteristics like textures, colours, and spatial gradients, guiding the generators to produce outputs with more pronounced matching-relevant characteristics. The multi-scale discriminator output fosters consistency at both global and local scales, enhancing overall performance. Additionally, using gradients directly as input proves advantageous compared to magnitudes or orientations, possibly due to neural network limitations in modelling orientation concepts effectively.

5 Conclusion and Future Works

The paper introduces ToDayGAN, a system designed for visual localization that utilizes image-to-image translation to enhance matching between night-time query images and a set of daytime reference images with known poses. A significant innovation lies in modifying discriminators within the CycleGAN/ComboGAN architecture, focusing on capturing low-level characteristics crucial for feature-based image matching. Specifically, ToDayGAN employs separate discriminators specialized in colour, luminance, and spatial gradients instead of a single discriminator, enabling a more nuanced analysis of input images. Additionally, these discriminators produce multi-scale outputs to promote consistency at both global and local scales.

The effectiveness of these modifications is demonstrated through extensive experiments conducted on the challenging Oxford RobotCar dataset. ToDayGAN achieves remarkable results, surpassing the previous state-of-the-art DenseVLAD method by over 2.5 times in accuracy on the standard 5m 10° pose threshold for night-time queries. Notably, it exhibits strong generalization capabilities, performing well even in scenarios like night rain, which significantly differs in appearance from the training data.

The authors draw several key conclusions from their ablation studies. Partitioning discriminators to focus on distinct low-level properties such as colour, texture, and gradients yields substantial performance improvements compared to using a single discriminator or concatenating inputs. Furthermore, employing spatial gradients as discriminator input rather than gradient magnitudes/orientations proves more compelling, possibly due to neural networks' challenges in modelling orientation concepts accurately. Including the relativistic discriminator loss enhances training stability compared to the standard adversarial loss. Additionally, post-processing techniques like dual evaluation contribute to further accuracy gains.

In terms of future work, the authors suggest several promising research directions. Exploring discriminator architectures tailored to different perceptual characteristics could enhance image quality and task performance across various generative tasks beyond image translation. Investigating alternative losses or constraints beyond adversarial and cycle consistency could enhance the reliability and consistency of generated images for specific applications. Extending the multi-scale discriminator output concept to other generative models may facilitate capturing image statistics at multiple scales. Moreover, applying techniques like ToDayGAN to other robotics/vision tasks, such as semantic segmentation or object detection, where addressing appearance gaps is crucial, holds significant potential. While the current work focuses on day-night scenarios, similar approaches could be extended to handle other dramatic appearance shifts like varying weather conditions or sensor modalities.

6 References

[1] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, and B. A. Research, “Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks.” pp. 2223–2232, 2017. Accessed: Apr. 19, 2024. [Online]. Available: https://github.com/junyanz/CycleGAN.

[2] A. Anoosheh and E. Agustsson, “ComboGAN: Unrestrained Scalability for Image Domain Translation.”

[3] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool, “Night-to-day image translation for retrieval-based localization,” *Proc IEEE Int Conf Robot Autom*, vol. 2019-May, pp. 5958–5964, May 2019, doi: 10.1109/ICRA.2019.8794387.

[4] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 year, 1000 km: The Oxford RobotCar dataset,” *http://dx.doi.org/10.1177/0278364916679498*, vol. 36, no. 1, pp. 3–15, Nov. 2016, doi: 10.1177/0278364916679498.