ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE
AND ENGINEERING

ARTIFICIAL INTELLIGENCE

MASTER THESIS

in

Image Processing and Computer Vision

NIGHT2DAY - AI ENHANCEMENT OF NIGHT PHOTOGRAPHY

CANDIDATE	SUPERVISOR
Davide Perozzi	Prof. Luigi Di Stefano

Academic year 2022-2023

Session 3rd

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Night photography often poses challenges due to reduced light availability, resulting in images with compromised details and colours. This can pose difficulty in the usability of this images across a wide range of applications (e.g. object detection, segmentation, environment description…) or just for qualitative human judgement. In response to this, artificial intelligence (AI) algorithms designed for images can come in hand leveraging their ability of features encoding and domain translation to be able to generate a day version of a night photo. This could enlarge the usability domain of the image, not just technical but also artistic.

We can just think about security cameras where a day conversion of the images would certainly help a lot the operator job without the need of using a more expensive night vision camera.

In this paper, we will describe our work revolved around the implementation of an AI algorithm designed for images translation able to receive as input a night image and convert it to the counterpart daytime version while maintaining the same context and structure of the original input.

The entire work has been conducted at the premises of CYENS - Centre of Excellence in Nicosia, Cyprus under the supervision of Dr. Alessandro Artusi, Xenios Milidonis and the help of the whole DeepCamera team.

In this paper we will illustrate the carried out work going through all the phases

from the objectives definition to the results evaluation.

Chapter 2

Objectives

Our initial objective was to build an AI algorithm able to receive as input a night image and output the daytime version maintaining the same context and structure of the original input.

The system could have been developed in-house but, due to the limitation in available time, we focused on leveraging open-source model architectures provided by the research community.

The main pillar of our objectives was to have a good generalization ability of the model to gain good performance for a wide range of images domains, since the models we encountered were all specialized on car dash cam view images. So, one of our goal was to collect a big amount of high quality images belonging to different domains.

Our performance goal was to maintain the metric scores of base model provided by literature (more in Chapter 5) and to get as close as possible to the visual performance of BrighterAI work shown by NVIDIA in their blog post ”Brighter AI Uses Deep Learning to Shed Light on Nighttime Video Footage” [3].

In general we aimed to reduce to the most the loss in context, details and information due to the translation and enhance generalization ability over more domains.

Chapter 3

Methodology

Our work started structuring a general timeline that gone through the months we had at disposal to give a pace to the work and deadline to be sure to optimize and do not waste time.

We started with a deep search in the literature of AI systems of night images enhancement, we have gone through several approaches comparing their specific peculiarities and performance.

At the same time, we were conducting a parallel investigation on available datasets of nighttime and daytime images, paired and unpaired ones. These phases gave us the clarity to choose the right approach for our case (discussed in details in Chapter 5).

Stated this, we collected datasets of different image domains to be able to give a wide representation of images distribution. We tried to collect the most data available possible while maintaining the higher quality standard. For our experiments we had at our disposal a NVIDIA GPUs cluster provided by the Municipality of Nicosia.

The months spent on this work were punctuated by weekly or bi-weekly meetings where problems, new ideas and next step to take were analyzed and discussed.

Chapter 4

Literature Review

The public literature proposes several approaches for the night to day task but we judged promising two main task domains and we focused on them: low-light image enhancement and image translation systems.

For the former the is a huge variety of architectures and models. The SOTA includes a huge variety of models, we analysed equalizers like Zero-DCE [5] (#1 on DICM and MEF datasets), encoder-decoder based architectures like LLFLOW [17] (#1 on LOL), GAN based architectures like EnglightenGAN [7] (#2 on DICM and MEF), transformer based architectures like LLFormer [16]. Despite this, all these proposals were not applicable to our case since we need a model able to fix the lights and color tones but also to adapt the output where information were missing (e.g. dark areas). Also a big obstacle we will talk about later is the lack of paired nigh to day datasets. These arguments leaded us to direct our focus to a more generative approach.

Image translation is a tasks domain more coherent with our goals. For this category there is not a proper SOTA since part of the metrics used to asses model performance are subjectives (e.g. visual evaluation).

We investigated Image-to-Image Diffusion Models like ControlNet [20], Palette [12] and BBDM [9], but despite the promising performance and the exhibition of remarkable adaptability to various use cases, we were not convinced by the immature technology and the reliance on paired datasets.

For these reasons, we focused on the more mature and battle tested systems proposed by the literature for Image-to-Image translation: basically were all a version of GAN or CycleGAN architectures.

ToDayGAN(2018) [2] is a ComboGAN-based model specifically trained for translation of images from night to day. It uses ComboGAN as the base image-to-image translation model, which is equivalent to CycleGAN in the case of two domains only.

AU-GAN(2021) [8] (Asymmetric and Uncertainty-aware GAN) is a CycleGAN-based model able to translate images from night to day and is particularly adapt to handle adverse weather conditions. This proposal is reported to perform better than base CycleGAN and toDayGAN and moreover it is technically more prone to adapt to a more various input domain, like we want.

Chapter 5

Implementation

The literature review described in Chapter 4 and data research described in Chapter 6 gave us the clarity that the best approach to choose was the unsupervised learning over a GAN architecture, specifically we choose the CycleGAN-based AU-GAN architecture.
First of all, let's describe the CycleGAN system:
CycleGAN, or Cycle Generative Adversarial Network, is a deep learning architecture that, unlike traditional paired image translation methods that require labeled training data, can learn to translate images between two domains without the need for explicit correspondences between the input and output images. This makes it versatile for a wide range of applications, including style transfer, colorization, and object manipulation.
The CycleGAN architecture consists of two main components:

- Generator Networks: CycleGAN employs two generator networks, each responsible for translating images from one domain to the other. For instance, one generator might translate images from horses to zebras, while the other translates zebras to horses. It consists of an encoder-decoder architecture. The encoder extracts features from the input image and compresses them into a latent representation. The decoder then

reconstructs the image from the latent representation, but in the target domain.

- Discriminator Networks: Alongside the generators, there are two discriminator networks, each tasked with distinguishing between real and translated images in its respective domain. The discriminators provide feedback to the generators, helping them improve their translation accuracy. The discriminator consists of a convolutional neural network (CNN) architecture with several convolutional layers, followed by pooling layers and fully connected layers.

The key innovation of CycleGAN is in its incorporation of cycle consistency loss. This loss function ensures that images are translated consistently across both domains. In other words, if an image from Domain A is translated to Domain B and then back to Domain A, the final output should closely resemble the original image. This cyclic consistency constraint prevents the generators from producing unrealistic or distorted translations.
We have different losses for the generator and discriminator components.
The generator loss consists of two components:

- Adversarial Loss: This loss measures the generator's ability to fool the discriminator. It is calculated using a binary cross-entropy loss function, where the generator aims to minimize the loss while the discriminator aims to maximize it.

- Cycle Consistency Loss: This loss ensures that images are translated consistently across both domains. It is calculated by comparing the original image to the image obtained after translating it back to the original domain. This loss encourages the generator to produce translations that are consistent and realistic.

The overall generator loss is the sum of the adversarial loss and the cycle consistency loss.

The discriminator loss is also a binary cross-entropy loss function. The discriminator's objective is to maximize the loss, distinguishing real images from translated ones. This loss encourages the discriminator to become more adept at identifying real images and rejecting translated ones.
During training, the generator and discriminator networks are optimized alternately. The generator is updated to minimize its loss, while the discriminator is updated to maximize its loss. This adversarial training process helps both networks improve their performance, leading to more accurate and realistic image translations.
Like said before, AU-GAN is based on CycleGAN but has several peculiarities that let this implementation to be more aligned with our objectives (see Fig. 5.1).
These are the main differences with CycleGAN:

- Asymmetric Architecture: AU-GAN adopts an asymmetric architecture where only one generator (GN→D) is enhanced with a transfer network (T-net). This allows for better handling of imbalanced information between domains. The T-net helps eliminate artifacts and other unwanted effects caused by adverse conditions, still allowing for an accurate translation.

- Feature Matching Loss: AU-GAN uses a feature matching loss that penalizes differences between encoded features from input images and their translated counterparts using different encoders.

- Uncertainty-aware Cycle-Consistency Loss: To address regional uncertainty, AU-GAN proposes an uncertainty-aware cycle-consistency loss that considers uncertainties present in reconstructed images. Incorporating the confidence map into the reconstruction loss, areas with higher uncertainty are given less weight in determining how well an image has

been reconstructed. This allows for more accurate assessment of reconstruction quality and helps to mitigate potential issues caused by artifacts or missing details.

This model is reported to achieve FID score of **38.6** on BDD100K testset. For our experiments we used the pretrained version of this model provided by the author itself in his GitHub [6], pretrained on BDD100K dataset [19]. More on this in Chapter 9.

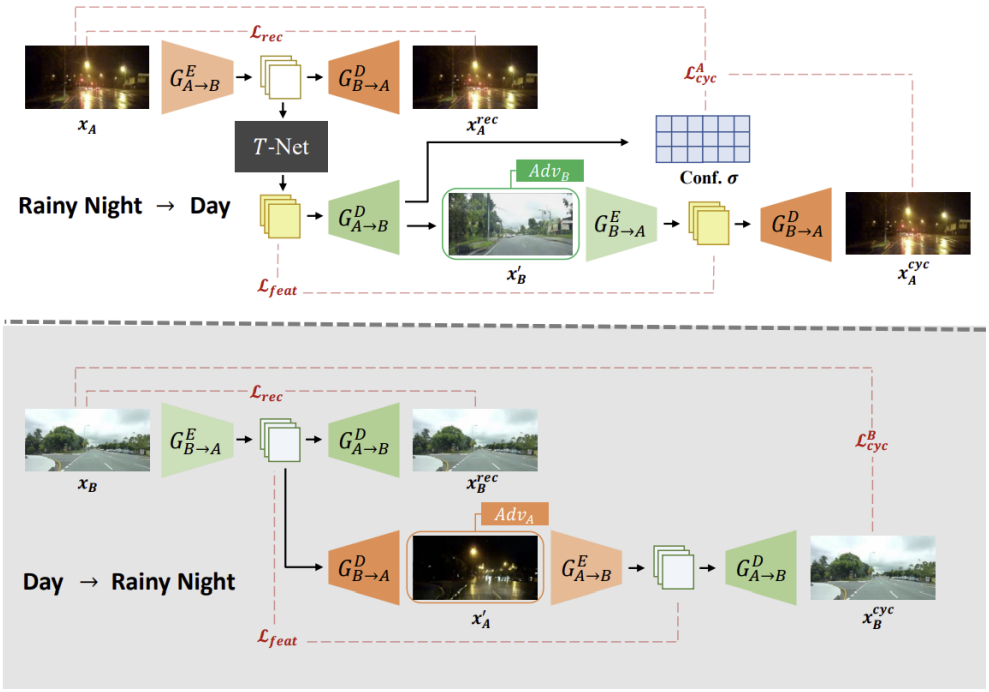

Figure 5.1: AU-GAN architecture

Chapter 6

Data Collection and Preprocessing

In parallel to the literature review, we were conducting an investigation on available datasets of nighttime and daytime images. We specifically searched for paired, unpaired and single domain dataset:

- Single domain dataset are the most common form of dataset because can be used for a huge variety of tasks. The problem using these datsets is the presence of unwanted images and it could take a lot of effort to filter them out for your needs.
- Unpaired dataset are formed by two sets of images respectively captured during night and day but are not coupled by any relation. The sets can also contain an uneven number of elements. Usually these kind of datasets have a common characteristic for the entire dataset (e.g. same city, same capture mode).
- Paired datasets are the most scarse type of datasets due to the difficulty on building them. For paired images we mean two images of the same exact subject (e.g. a building, a cityview) taken in the two versions night and day. It is fundamental that the two images have to be captured in the same position in order to have a nearly perfect matching of the pixels between the two versions. These kind of datasets are usefull for

a supervised learning approach but a lot of them are synthesized with algorithms starting from one of the two domain.

Based on our search there is a shortage of paired datasets and most of them are either synthetic (N2D250K[11], MIT-Adobe FiveK[10]) or contain too many images of domains we are not interested in and/or are taken with low exposure instead of during real night (LOL[18], SID [4]). A real day and night paired dataset is the Day-night Dataset but the quality of the images is too low. The unpaired category is the most promising one and we selected the following datasets based on image domain quality and size:

- Unpaired Day and Night cityview images [15], 549 images, panoramic cityview images (see Fig. 6.1)

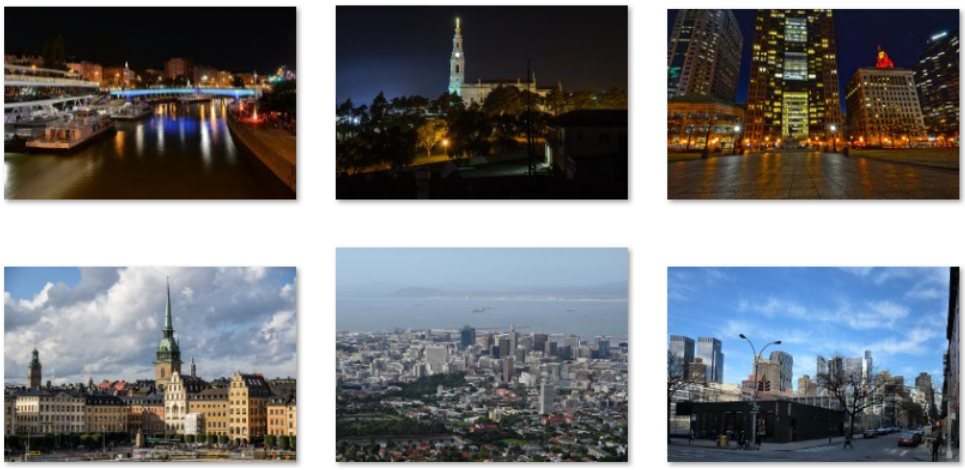

Figure 6.1: Unpaired Day and Night dataset samples

- 24/7 Tokyo [14], 1125 images, perspective of a pedestrian in a city (see Fig. 6.2)
- Aachen Day-Night [1], 3401 images, perspective of a pedestrian in the city (see Fig. 6.3)
- Dark Zurich [13], 8779 images, perspective of a car dash cam (see Fig. 6.4)

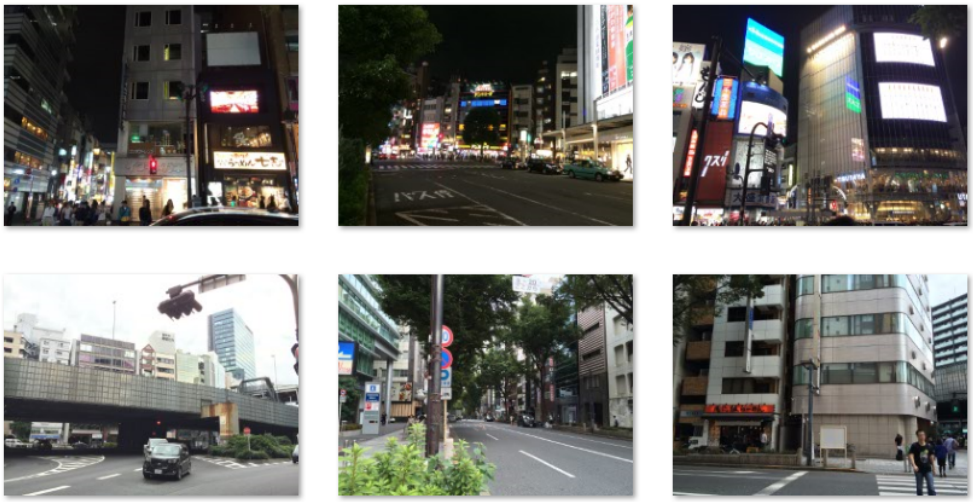

Figure 6.2: 24/7 Tokyo dataset samples

Figure 6.3: Aachen Day-Night dataset samples

These research gave us the clarity that the best approach to choose was the unsupervised learning, since the shortage of paired datasets were difficult to handle and we had not the time and resources to build a new one from scratch.

For the preprocessing phase we opted for sampling to reduce size, removal of unwanted category and resizing to normalize dimensions.

Dark Zurich is a big dataset that is built from frames of videos, so there are a lot of redundant images since lot of the sequential images are nearly the same. To address this, we sampled 1 image each 4 ending up with a subset of 1/4 of

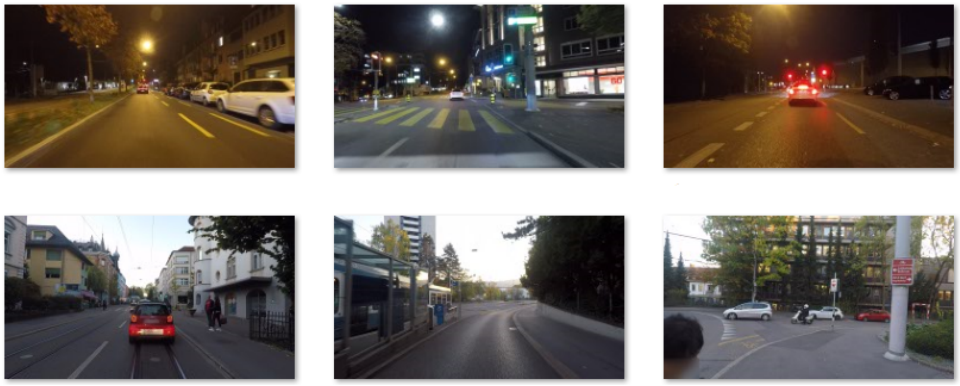

Figure 6.4: Dark Zurich dataset samples

the original size without compromising the general information content of the dataset.

For 24/7 Tokyo dataset we removed all the images taken in sunset since we wanted only a binary dataset of night and day images.

Aachen Day-Night and Unpaired Day Nigh datasets were untouched.

We ended up with the number of elements shown in Table 6.1:

dataset/category	**Day**	**Night**	**Total**
24/7 Tokyo	375	375	750
Aachen	833	206	1039
Unpaired Day Night	522	227	749
Dark Zurich	760	1334	2094

Table 6.1: Datasets images distribution.

After the preprocessing, we merged all the datasets to form a general big dataset of 4632 images (2490 day images + 2142 night images) called **Omniset**. We merged together the datasets with same domain: **247 Tokyo** and **Aachen Day-Night**.

For testing purposes, we randomly extrapolated a small testset of 5 images from each single datasets (20 total). To maximize the size of the training set, we deliberately opted for a minimal number of elements in the test set. While we acknowledge that the test set's size can influence the FID score (more on FID score in Chapter 8), our primary focus was to assign greater significance

to human evaluation.

Chapter 7

Experiments

After the collection of the data we needed, we started with the training phase during which we conducted several experiments.
Specifically:

- **Single-domain Continue Training**:
  We trained different instances of the model over each of the four single-domain datasets. We wanted to asses the impact on performance of each domain. We performed a Continue Training since the model provided by the author was already pretrained over BDD100K dataset.

- **Multi-domain Continue Training**:
  We trained the model over the Omniset dataset.

- **Multi-domain Continue Training - Uncertainty Awareness Disabled**:
  We trained the model over the Omniset disabling the uncertainty awareness. We decided to perform this experiment due to the results of the previous experiments. More on this in Chapter 9.

- **Multi-domain Fine-tuning - Generator Encoder Freezed**:
  We trained the model over Omniset without updating the Generator Encoder weights. We wanted to asses the learning capacity of the architecture freezing this single component, letting him to use his prelearned

capacity of encoding the features of the input image.

- **Multi-domain Fine-tuning - Generator Encoder Singular Layer Freezed**: We trained different instances of the model over Omniset freezing for each one a different layer of the Generator Encoder. We wanted to analyze the importance of each layer in the encoding capacity of the model.

The hyper-parameters we used for each training are:
epochs: 5,
batch size: 8,
input image height: 256.
Such a low number of epochs was mandatory due to the enormous amount of time needed to train, near 10 days for a 5 epochs session. Despite we had at our disposal a GPU cluster, the training code provided by the author was not compatible with a multi-GPU training. While adapting the code for multi-GPU usage could have reduced the training time, unfortunately we had no time to invest in a code conversion so we opted for a single-GPU training.

Chapter 8

Results

Starting from the training, we can clearly see a slight descent of the losses and then the appearance of a plateau (see Fig. 8.1). We can not determine if this

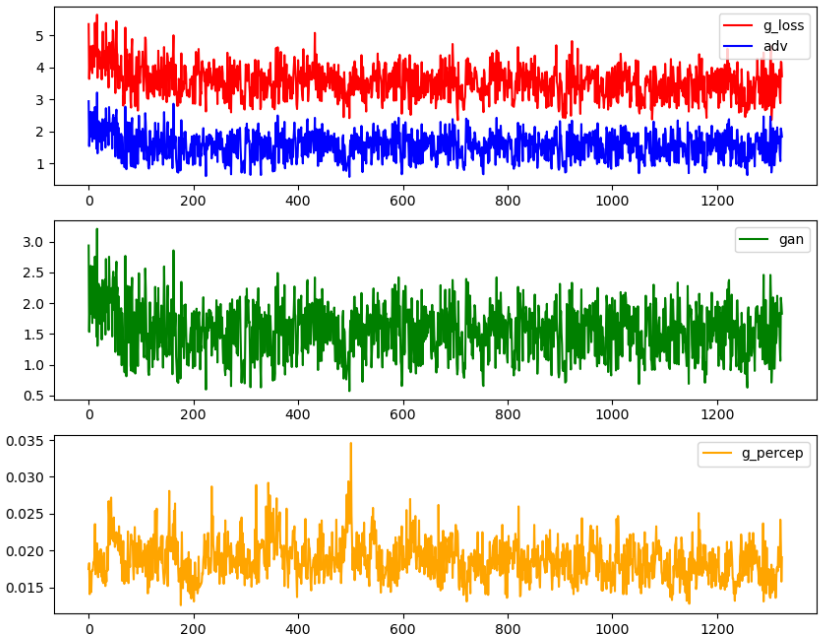

Figure 8.1: Plot of training metrics. Batches processed as *x axis*

is a local plateau or a global one since the number of epochs is very small and can not be enlarged due to the training duration.

We performed both quantitative and qualitative tests on our trained models.

For the former, we choose to use the Fréchet inception distance (FID) metric:

$$FID(x, g) = ||\mu_x - \mu_g||^2 + Tr(\Sigma_x + \Sigma_g - 2(\Sigma_x\Sigma_g)^{1/2}) \tag{8.1}$$

The Fréchet Inception Distance (FID) is a metric used to assess the quality of generated images produced by generative models, like GANs. FID is based on feature vectors extracted from a pre-trained neural network (Inception in our case). It measures the similarity between the distributions of feature vectors for real and generated images. Lower FID values indicate better image quality.

For the qualitative tests, we conducted an internal human judgment evaluation with our team.

Both quantitative and qualitative tests have been conducted using:

- single-domain testsets, to assess granularity of performance in each domain.
- omniset testset (single-domain testsets merged), to assess generalization ability.
- 1000 images subset of BDD100K testset, to compare performance with the pre-trained model in the paper.

The models' name in tables refers to the domain over which they have been trained.

- **pre-t** refers the pretrained model provided by the author,
- **pedestrian** refers the model trained over 247Tokyo and Aachen Day-Night,

- **dashcam** refers the model trained over Dark Zurich,
- **cityview** refers the model trained over Unpaired Day and Night,
- **omniset** refers to the model fine-tuned over Omniset,
- **omniset EF** refers to the model fine-tuned freezing the encoder layers of the generator,
- **omniset -UA** refers to the model with confidence marix disabled,
- **omniset L*n*** refers to the model with layer *n* of the encoder freezed during training.

The results of the FID metric are the following:

testset / model	**pre-t**	**pedestrian**	**dashcam**	**cityview**	**omniset**
omniset	581.68	577.03	576.26	575.13	**509.48**
BDD100K	220.52	218.95	218.19	220.24	**188.44**
247tokio	673.73	628.13	635.68	647.26	**511.10**
darkzurich	548.63	544.73	543.74	**541.79**	615.42
unpaired day night	675.62	671.79	665.26	688.64	**631.53**
aachen	505.32	521.78	538.19	516.05	**411.07**

Table 8.1: Comparison of models trained on single-domain datasets and multi-domain dataset.

testset / model	**omniset**	**omniset EF**
omniset	**509.48**	534.39
BDD100K	**188.44**	191.16
247tokio	**511.10**	569.00
darkzurich	615.42	**552.10**
unpaired day night	**631.53**	642.09
aachen	**411.07**	425.63

Table 8.2: Comparison of models trained on multi-domain dataset with and without encoder freezed.

testset model	omniset	omniset -UA	omniset L1	omniset L2	omniset L3
omniset	**509.48**	532.55	555.82	558.56	591.43

Table 8.3: Comparison of models trained on multi-domain dataset with different layers freezed.

Regarding the qualitative test, in Figure 8.2 are shown some good results of the model full-trained on Omniset:

Figure 8.2: Samples of qualitative good results.

In Figure 8.3 are shown some bad results:

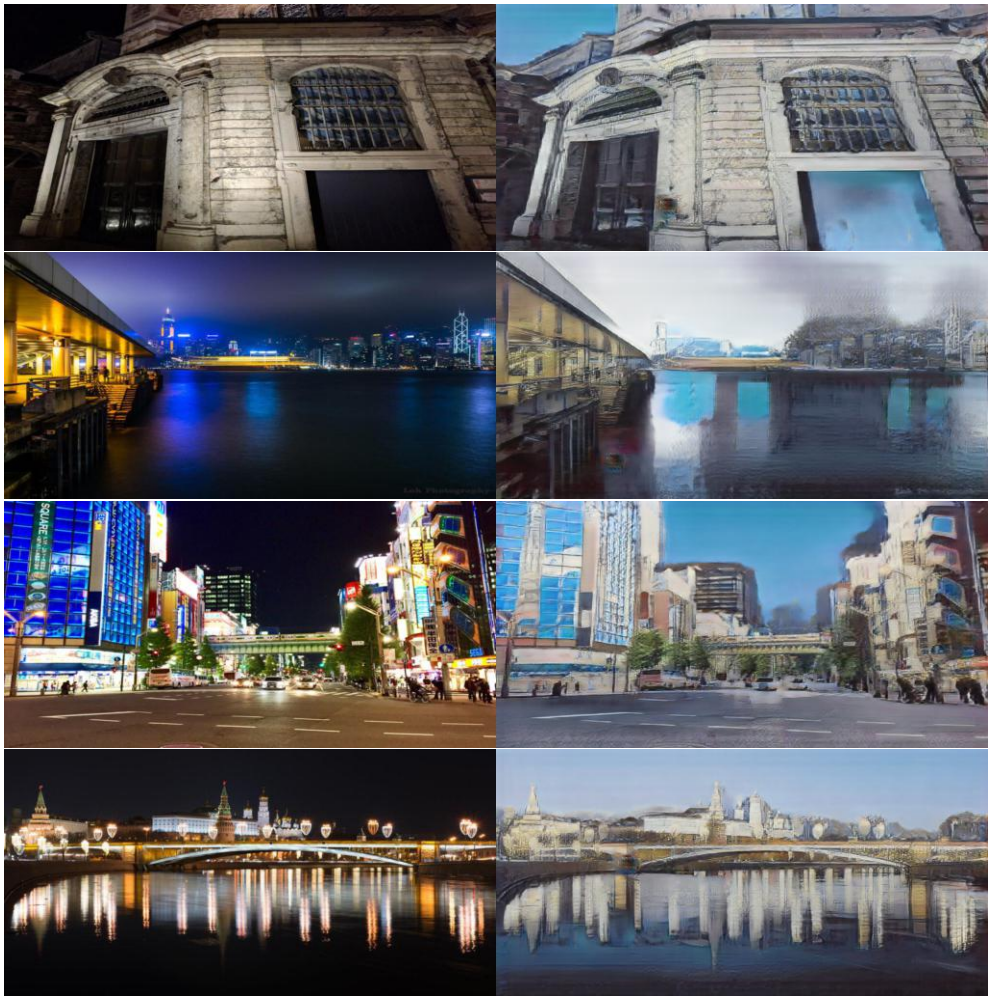

Figure 8.3: Samples of qualitative bad results.

Chapter 9

Discussion

The results confirmed the behaviour we were expecting. The training over a wider domains range enhanced the generalization ability of the model; this is clearly enlightened by the superior performance of the multi-domain trained model against the pre-trained base model and the single-domain finetuned models on any testset (except for Dark Zurich testset).
The superior performance of the model with the freezed encoder against full-trained model on Dark Zurich testset we suppose can be due to the pretraining on the same image domain (BDD100K dataset, car dash cam perspective), this could have played a relevant role in enhancing the ability of handling the context of that kind of domain.
Also the experiment on freezing single layer gave us some insights on pre-learned capability. The fine-tuning freezing single layers determined a decrease of FID performance with the freezing of lower layers (L2 and L3), this could indicate a L1's stronger role in feature extraction since the model perform better when it is kept as pre-trained instead of L2 or L3. Despite this, updating all the layers is confirmed to bring the best results. Moreover, The ablation study of the confidence matrix confirmed its role in enhancing the performance when enabled, since the FID score decrease when it is not used.

Despite these findings, the model performance we ended up with are not satisfactory. The generalization ability is effectively enchanced but the best

visual performance are still with the domain used for pre-training, car dash cam view. The model struggles on images with bright and saturated colours having difficulty in handling tonality, also it has difficulty with light reflections (e.g., on water) and some large black regions (i.e. it fills them creatively or confusing them for the sky). Also, some blurry effects and distortions are noticeable in regular pattern zones (e.g. windows, bricks). We though this could be due to the confidence matrix but the previously shown results confuted our idea.

The probable reasons of these poor results are to address to the small dimension of the datasets used but the most to the original performance of the base model.

In that regard, we strongly believe that the pre-trained model provided by the author in his GitHub is not the same mentioned and tested in the presentation paper, since the results we have with it are orders of magnitude worse than the ones shown in the paper.

Chapter 10

Conclusions and Recommendations

This work gave us the opportunity to build a system from the idea to the actual testing phase in a real professional environment. Also we were able to put the hands on a big model like a CycleGAN and asses its powers but also its drawbacks. In these months of developing a lot of work have been done but a lot more could be done.
There a several improvements that can be implemented in future works.

- We suggest to enlarge the dataset with way more samples from the wider range of domains possible while maintaining high the quality of the images, the size and the quality of the dataset could be the main factors to achieve satisfying results.
- Retrain the base model from scratch with the new enlarged dataset, this should give the model a good generalization ability from the begin.
- Fine-tune the base model on some specific domain based on the future model application.
- Adapt the code to be able to train on multi-GPU in order to increase the

number of epochs and asses if the training metrics plateau is local or general.

Bibliography

[1] Aachen Day-Night datasets. URL: https://www.visuallocalization.net/datasets/.

[2] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. V. Gool. Night-to-day image translation for retrieval-based localization. *arXiv preprint arXiv:1809.09767v2 [cs.CV]*, 2019.

[3] Brighter AI Uses Deep Learning to Shed Light on Nighttime Video Footage. URL: https://blogs.nvidia.com.tw/2017/11/30/brighter-ai/.

[4] C. Chen, Q. Chen, J. Xu, and V. Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[6] jgkwak95's AU-GAN GitHub. URL: https://github.com/jgkwak95/AU-GAN.

[7] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang. Enlightengan: deep light enhancement without paired supervision, 2021. arXiv: 1906.06972 [cs.CV].

[8] J.-g. Kwak, Y. Jin, Y. Li, D. Yoon, D. Kim, and H. Ko. Adverse weather image translation with asymmetric and uncertainty-aware gan (au-gan). In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021.

[9] B. Li, K. Xue, B. Liu, and Y.-K. Lai. Bbdm: image-to-image translation with brownian bridge diffusion models, 2023. arXiv: 2205.07680 [cs.CV].

[10] MIT-Adobe FiveK Dataset. URL: https://data.csail.mit.edu/graphics/fivek/.

[11] N2D250K dataset. URL: https://github.com/isurushanaka/N2D250K.

[12] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi. Palette: image-to-image diffusion models, 2022. arXiv: 2111.05826 [cs.CV].

[13] C. Sakaridis, D. Dai, and L. V. Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV)*. ETH Zurich and KU Leuven, 2019.

[14] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, 2015.

[15] Unpaired Day and Night cityview images. URL: https://www.kaggle.com/datasets/heonh0/daynight-cityview/data.

[16] T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, and T. Lu. Ultra-high-definition low-light image enhancement: a benchmark and transformer-based method. *arXiv preprint arXiv:2212.11548v1 [cs.CV]*, December 2022.

[17] Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, and A. C. Kot. Low-light image enhancement with normalizing flow. *arXiv preprint arXiv:2109.05923v1 [eess.IV]*, 2021.

[18] C. Wei, W. Wang, W. Yang, and J. Liu. Deep retinex decomposition for low-light enhancement, 2018. arXiv: 1808.04560 [cs.CV].

[19] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: a diverse driving dataset for heterogeneous multitask learning, 2020. arXiv: 1805.04687 [cs.CV].

[20] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023. arXiv: 2302.05543 [cs.CV].

Acknowledgements

I would like to express my gratitude to the University of Bologna and CYENS - Centre of Excellence for the opportunities they gave me and to Professor Luigi Di Stefano, Dr. Alessandro Artusi, and the entire DeepCamera team for their guidance through my research journey.

Inexpressible gratitude is reserved for my family, whose support has been fundamental for my entire journey.

A golden mention is deserved for my life companion, Arianna. She has been the pillar of my existence. Without her, nothing would have been the same.